TECHNICAL MEMORANDUM

438

Report on the sixteenth meeting of Computing Representatives 21–22 April 2004

P. Prior (Compiler)

Operations Department

October 2004

European Centre for Medium-Range Weather Forecasts
Europäisches Zentrum für mittelfristige Wettervorhersage
Centre européen pour les prévisions météorologiques à moyen terme

Series: Technical Memoranda

A full list of ECMWF Publications can be found on our web site under:
http://www.ecmwf.int/publications/

Contact: library@ecmwf.int

European Centre for Medium Range Weather Forecasts
Shinfield Park, Reading, RG2 9AX, England

Contents

Preface

The sixteenth meeting of Computing Representatives took place on 21-22 April 2004 at ECMWF. Eighteen Member States and Co-operating States, plus EUMETSAT, were represented. The list of attendees is given in annex 1.

The Head of the Computer Division (Walter Zwieflhofer) opened the meeting and welcomed representatives. He gave a presentation on the current status of ECMWF's computer service and plans for its development. Each Computing Representative then gave a short presentation on their service and the use their staff make of ECMWF's computer facilities. Participants were also invited to report on their experience in operating and managing Linux clusters, to complement the assessments ECMWF were planning to carry out as background knowledge for future HPCF replacements. There were also presentations from ECMWF staff members on various specific developments in the ECMWF systems. The full programme is given in Annex 2.

This report summarises each presentation. Part I contains ECMWF's contributions and general discussions. Part II contains presentations on Linux experiences and Part III Member States' and Co-operating States' contributions; all the reports were provided by the representatives themselves.

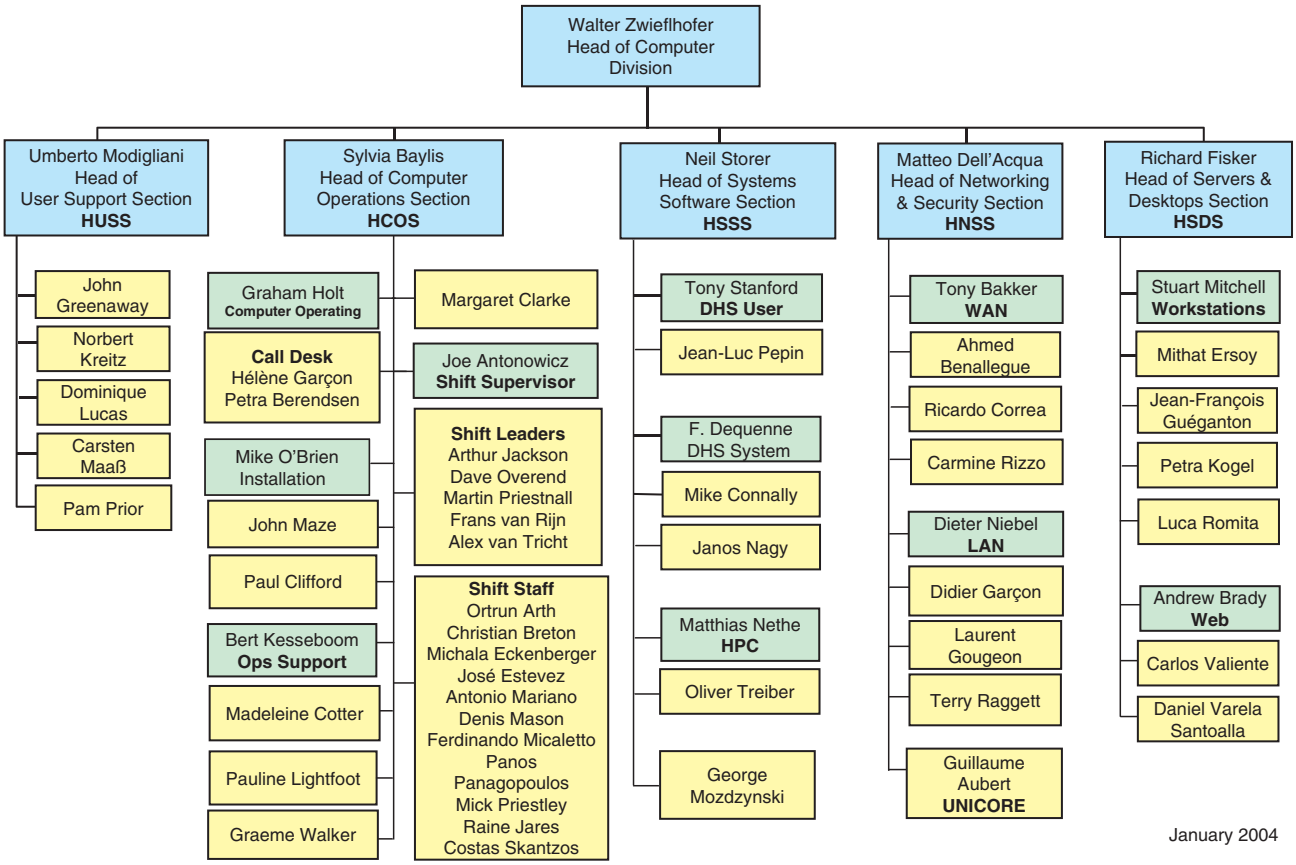
Part I

ECMWF Staff contributions and general discussions

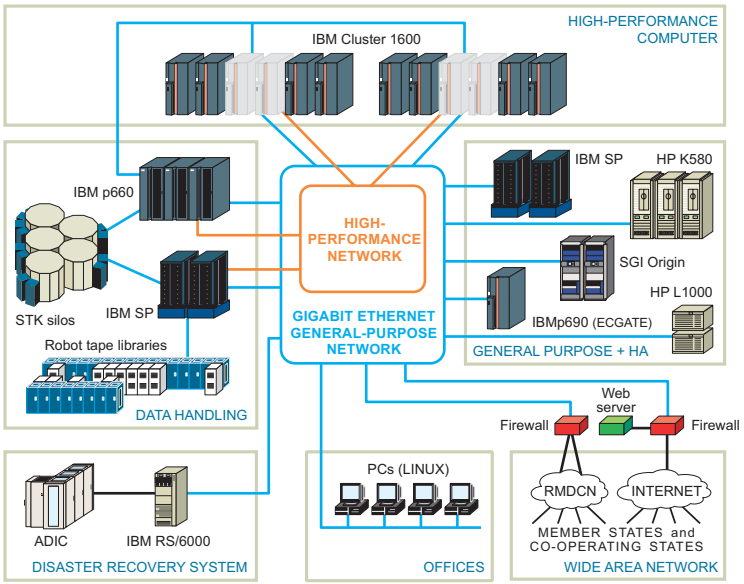
ECMWF Computing Service: Status and Plans - *Walter Zwieflhofer, Head of Computer Division*

Major activities over the past 12 months

- Phase 1 of the IBM HPCF continues to provide an excellent service at a high level of availability
- Improvements to HPCF job scheduling were made
- A new chiller and air handling units were installed to provide additional cooling capacity for HPCF Phase 3
- Phase 3 of the new DHS is being installed
  - MARS has been completely migrated
  - ECFS migration started in February 2004
- New IBM server ecgate was installed last year
- New Entity Management System has been implemented
- RMDCN upgrade of the Base Package was completed
- The review of Computer Operations Section was completed and recommendations are being implemented
- A High-Sensitivity Smoke Detection system was installed



Computer Division Organigramme



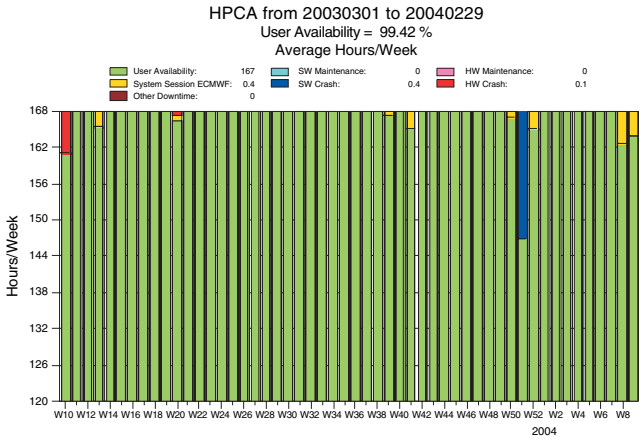
ECMWF Computer Environment

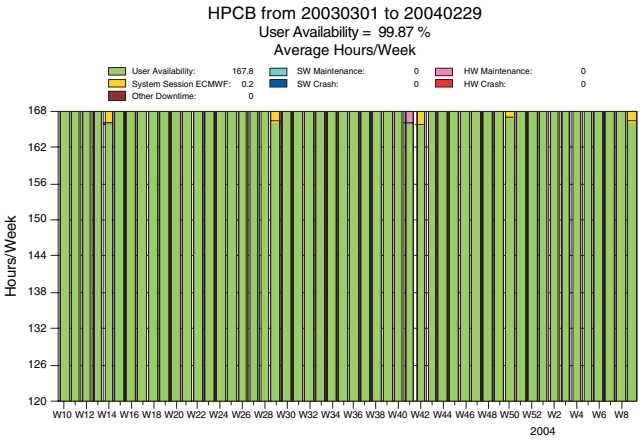
Phase 3 of the IBM HPCF

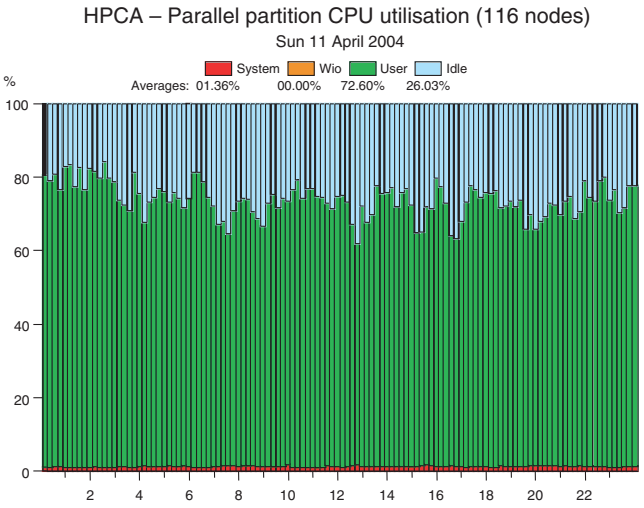
- Two identical clusters with 68 p690++ servers each
- Nodes for user work:
  - Each cluster has 66 32-processor servers for user work
  - 6 of the servers in each cluster have 128 GB memory
  - All other nodes have 32 GB memory
  - Processors (Power4+) run at 1.9 GHz (7.6 Gigaflops peak)
  - ~25 terabytes of disk per cluster
  - p-Series High Performance Switch (4 links per server)
- Nodes for I/O and networking
  - 2 p690++ servers, each partitioned into several smaller nodes
- The number of nodes is an estimate and could go up or down depending on the results of the performance test

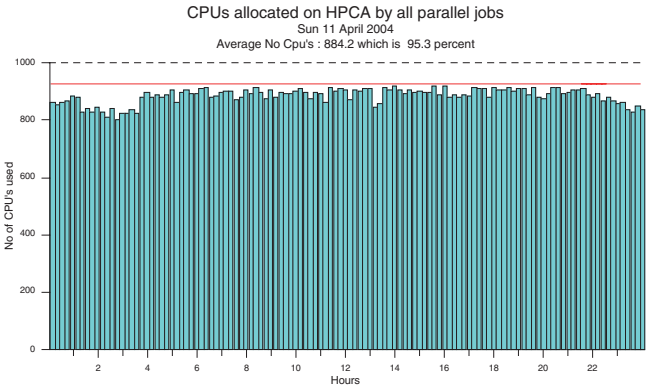
HPC Phase 3 schedule

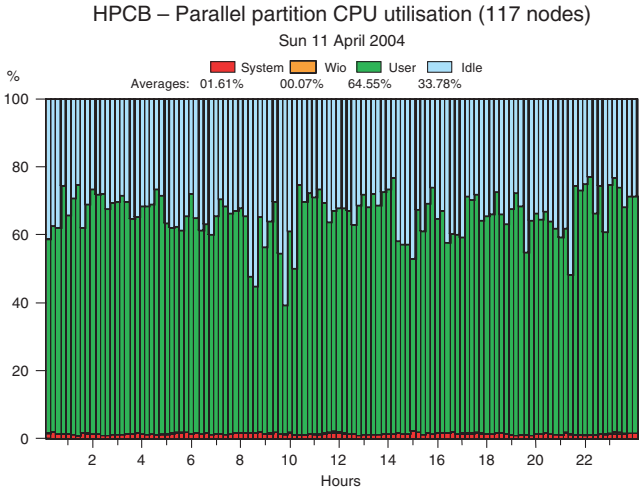
March	Start to build Phase-3C in Poughkeepsie
April	Start to build Phase-3D at ECMWF
May	Configure, test and set up to run Acceptance Tests on 3D
June 23	Start of acceptance of 3D (including Operational Test)
July	Allow Member State users access to 3D to prepare for migration - manpower resources are available to assist these users Move the Operational Suite to 3D
August	Decommission 1B Start to build and test 3C at ECMWF
September	Decommission 1A, complete the build of 3C
October	Configure, test and set up to run Acceptance Tests on 3C
October 26	Start acceptance of 3C
January 2005	Complete Acceptance of the whole Phase 3 system

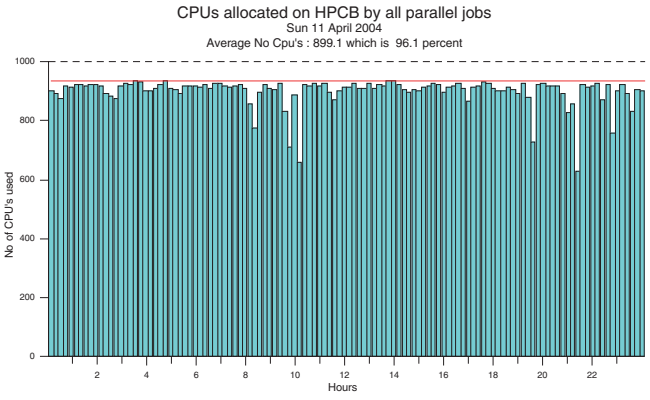

HPCA from 20030301 to 20040229
User Availability = 99.42 %
Average Hours/Week
User Availability: 167
System Session ECMWF: 0.4
Other Downtime: 0
SW Maintenance: 0
SW Crash: 0.4
HW Maintenance: 0
HW Crash: 0.1
Hours/Week
2004


HPCB from 20030301 to 20040229
User Availability = 99.87 %
Average Hours/Week
User Availability: 167.8
System Session ECMWF: 0.2
Other Downtime: 0
SW Maintenance: 0
SW Crash: 0
HW Maintenance: 0
HW Crash: 0
Hours/Week
2004

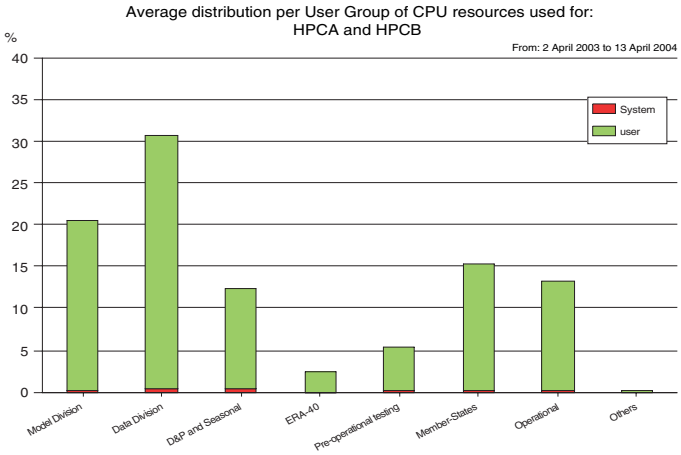

HPCA – Parallel partition CPU utilisation (116 nodes)
Sun 11 April 2004
System Wio User Idle
Averages: 01.36% 00.00% 72.60% 26.03%
%


CPUs allocated on HPCA by all parallel jobs
Sun 11 April 2004
Average No Cpu's : 884.2 which is 95.3 percent
No of CPU's used
Hours


HPCB – Parallel partition CPU utilisation (117 nodes)
Sun 11 April 2004
System Wio User Idle
Averages: 01.61% 00.07% 64.55% 33.78%
%
Hours


CPUs allocated on HPCB by all parallel jobs
Sun 11 April 2004
Average No Cpu's : 899.1 which is 96.1 percent
No of CPU's used
Hours

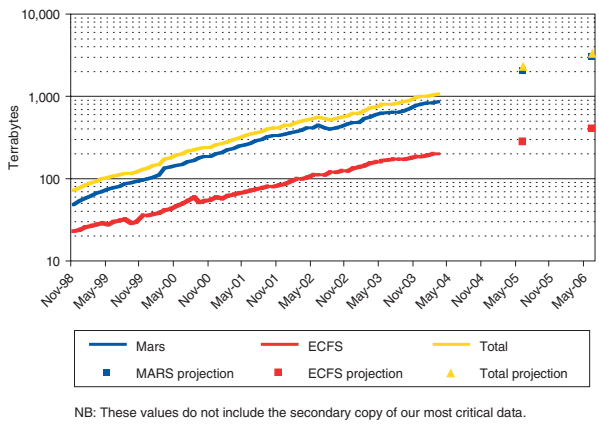
Distribution of HPCA and HPCB resources per user group



DHS

- The new HPSS-based system is performing very well
- All of the Phase 2 equipment and most of the Phase 3 equipment has been installed
- The system consists of:
  - 4 IBM p650-6M2 servers
  - 2 IBM p660-6H1 servers
  - 1 IBM p660-6m1 server
  - 28 IBM 3590H tape drives
  - 60 IBM 3592 tape drives
  - ~28 TB of disk space
- MARS uses only HPSS
- Most new ECFS data is stored in HPSS and the back-archive of the TSM ECFS data is underway
- Last year we upgraded HPSS from version 4.5 to version 5.1. This was a major upgrade requiring a migration of the HPSS database to DB2 format. This was required before we could offer an ECFS service using HPSS
- The introduction of the new ECFS service is transparent to users - the new ECFS handles data stored in both the old and the new DHS
- With the HPSS version of ECFS no automatic secondary copy is made of ECFS data (as there was with the old ECFS system). The user has to specify the "-b" option on the "ecp" command to request a secondary copy to be made

Volume of data stored



NB: These values do not include the secondary copy of our most critical data.

Servers and desktops

- The Linux systems continue to be very stable
- An upgrade of the desktop systems to latest versions of the various system components (SUSE 9.1, KDE 3.2, VMWare 4, Windows XP, Office 2003, ...) is being planned
- The main internal SGI servers were shutdown as planned; one Origin 2000 still used by the ERA project
- ecgate1 was upgraded using some of the above SGI equipment in May 2003
- Replacement for ecgate1 chosen and installed last year
- Scratch disk space was increased to ~½ TB
- Available for trial service in early December 2003, full user service started on 22 January 2004.

Ecgate

- Ecgate is an IBM p690
  - 16 1.3 GHz Power 4 CPUs
  - 32 GB Memory
  - 1 TB disk subsystem (IBM FASTt700)
  - Running AIX 5.2 and using LoadLeveler as a batch system
- Very similar to the nodes comprising HPC Phase1 (but with only 16 CPUs instead of 32)
- The service on ecgate1 (SGI Origin) will continue until the end of July 2004
- Please encourage users to start migrating to ecgate as soon as possible
- Please refer to Umberto's presentation for more information and discussion on status of migration

Linux Cluster

- Linux Cluster will be installed in late April for evaluation
- Supplied by Linux Networx
- Configuration is
  - 32 nodes plus 1 master node
  - Includes 6 I/O nodes with Fibre Channel HBAs
  - Each node has dual 2.2 GHz AMD Opteron 248 CPUs, 4 GB Memory
  - InfiniBand low latency high bandwidth interconnect for MPI
- Plan to evaluate shared/parallel file systems, particularly Lustre
- Goals are to evaluate this technology both for future HPC requirements and for general purpose servers

Entity Management System

- The Entity Management System (EMS) has been implemented at ECWMF to replace the previous user registration and authentication system
- It can cope with the different types of users and organizations ECMWF deals with
- The core of the system became operational in December 2003 and is initially being used internally by the ECMWF Call Desk to register both internal & Member State users
- The system is being extended to enable Computing Representatives to carry out certain registration tasks directly via a browser interface
- The interface for Computing Representatives should be available by the Summer this year

Web services

- The ECMWF web servers continue to provide a stable and reliable service. New content includes:
  - ENACT and ERA40 data added to the Research Data Web Service
  - Forecast charts, increased parameters, 12UTC and 10 day archives
  - Library bibliography is now a database driven web application
  - Ecgate documentation
- The growth in use of the web site continues to increase:
  - Total number of page accesses in 2003 11 million
  - Average page accesses 1 every 3 seconds
  - Change compared with 2002 +35%
  - Total number of accesses by users 1.6 million
  - Change compared with 2002 +37%
  - Ratio of recognised to public users 1 in 7
- A considerable revision of the web login will be introduced soon :
  - Users will be able to have a persistent login (no more lost rooms);
  - Users will be able to use password, certificate or SecuridID
  - Certain pages may demand higher authentication (eg SecurID for PrepIFS);
  - Domain login continues but will not be sufficient for a Room;
  - Domain users will not be transparently logged in.
- A mailing list management system (Sympa) has been implemented and will be used to contact external users

RMDCN

- The upgrade of the Member States' Base Package was successfully completed in mid-March 2004. The Centre now has two 34 Mbps access lines to the RMDCN and PVCs to Member States and Co-operating States range from 64 Kbps to 768 Kbps
- New members:
  - Japan joined the RMDCN at end October 2003 and their connection to China was accepted in January 2004
  - India signed an Accession Agreement on 17 February 2004, Their connection to Tokyo and Moscow should be ready in early June
  - Serbia and Montenegro is being connected to the RMDCN
  - Luxembourg is being connected to the RMDCN
- RMDCN Price and Technology reviews have started.

First results are expected by early June.

ECaccess

- ECaccess portal was enhanced to provide access to the MARS archive. A release of Metview including the support of the ECaccess-based "ecmars" was made available in June 2003
- ECaccess has been enhanced to provide support for LoadLeveler job submision
- ECaccess gateways are now installed in the majority of MS/Co-operating states
- Connections to ecgate and HPCA are now possible via any ECaccess gateways
- The previous telnet, ftp and X11 gateway for access to ECMWF via Internet was terminated at the end of December 2003
- The services provided by the ecbatch/eccopy software will be terminated with the decommissioning of ecgate1

LAN

- Following last year's ITT, Force 10 equipment was selected for the replacement of the High Performance Network
- Two E600 switches interconnected by 2 10GE links form the core of the network
- Phase 1 was delivered on 1 March 2004 and is currently under acceptance. DHS and HPC system were connected to the Force10 switches mid-March
- Phase 2 will be delivered in September 2004 and will include new high density module.
  - the 2 core switches will be interconnected by 4 10 GE links
  - HPCF Phase 3 systems will be connected to the HPN with trunks of 4 GE links
- Wireless LAN was installed in the conference block at the end of 2003
  - Access to the Internet is offered to external people attending meetings and conferences
  - Requires userid and password

Infrastructure work

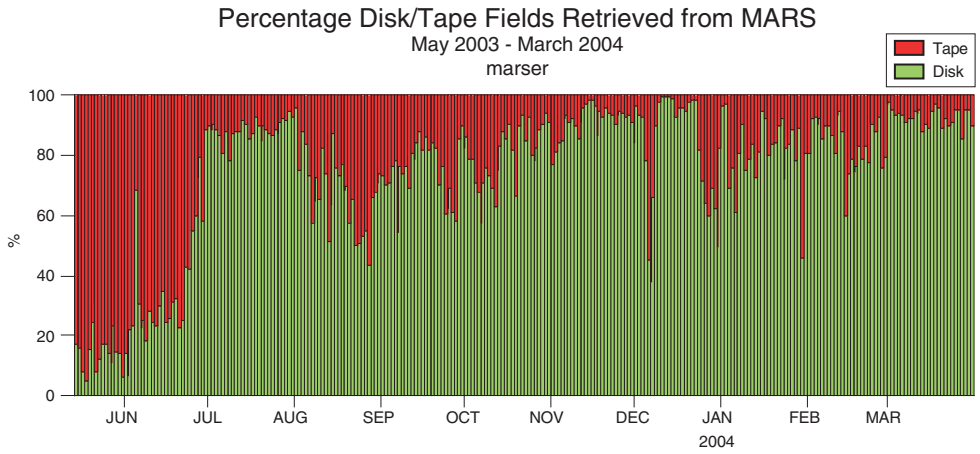
- An additional 11 kV supply was installed with a different cable route from the existing supply
- A new 2MVA Uninterruptible Power Supply system was ordered
  - to provide increased UPS capacity and restore N+1 resilience
  - to replace one of the old standby generators
  - the output can be split so it will run as UPS and standby generator
  - installation will be completed in summer 2004
- ITTs for the extension of the Computer Hall (building, electrical and mechanical services)
  - issued in February; expect to select the successful tenderer in May
  - building should be completed by September 2005
- Installation of High Sensitivity Smoke Detection

Other activities

- The current dissemination software, QFTD, has reached its limitation and is being re-developed. The new system, ECPDS, will offer different transport mechanisms (FTP, SFTP, gridFTP, ...) and the possibility of using the ECaccess network to disseminate securely over the Internet
- Involvement in 2 proposed EU projects:
  - DEISA includes a number of European supercomputing centres and aims at developing/operating a distributed super-cluster plus GRID-based interfaces to other large supercomputer sites
  - SIMDAT includes participants from aerospace, automotive, pharmacy and meteorology. One objective is to build a VGISC reference implementation for the future WMO Information System

Actions from the previous meeting

- About 4 TB of disk space was added to the new DHS system to cache a large part of the ERA40 archive. This has significantly improved the retrieval times for ERA40 data

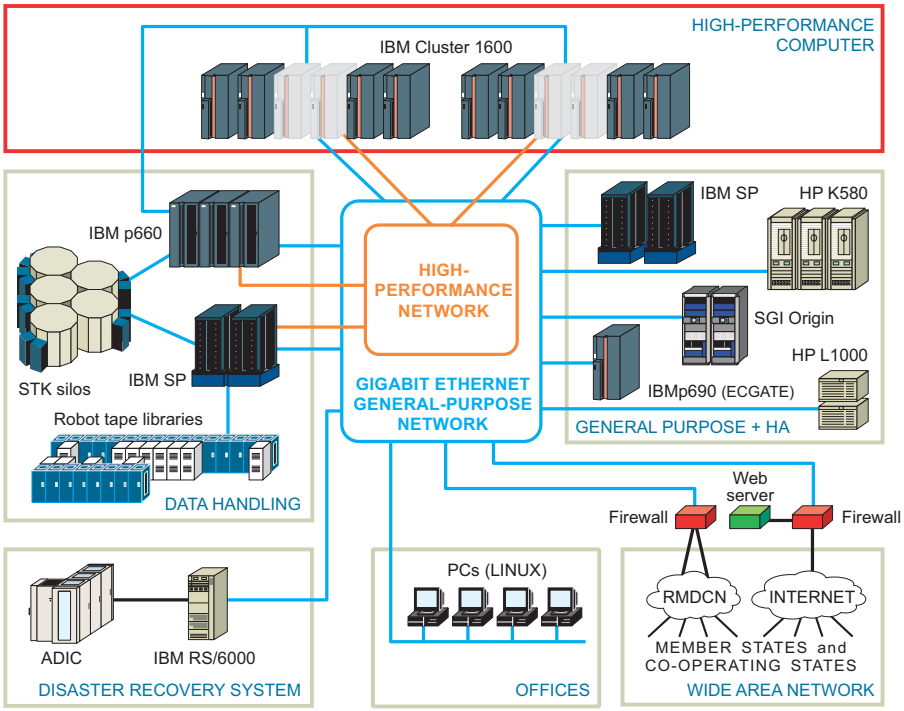


Major ongoing/planned activities

- Complete installation and acceptance of HPCF Phase 3
- Complete the migration of ECFS data from TSM to HPSS
- Decommission ecgate1
- Enable Computing Representatives to carry out certain registration tasks
- Perform RMDCN Price and Technology review
- ITT for the HA pre-processing and dissemination system
- Install Phase 2 of the high-performance LAN
- Complete the installation of UPS enhancement
- Installation of an inert gas fire suppression system
- Start the work on the extension of the Computer Hall

HPCF & DHS update - *Neil Storer, Head of Systems Software Section*

HPCF



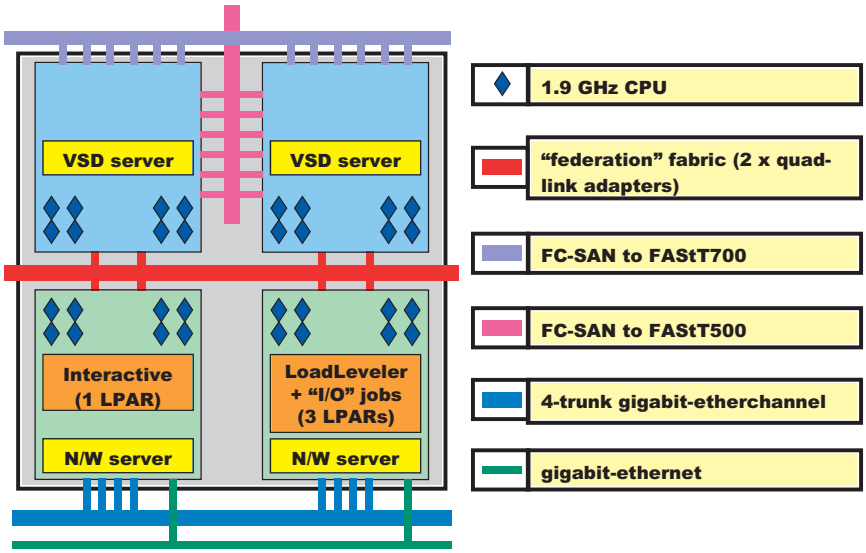
Phase 3 of the IBM HPCF

- On each of the Phase 1 clusters there are 30 p690 servers, each partitioned into 4 nodes (8-CPUs). The 66 or so Phase 3 compute servers will not be partitioned, so each Phase 3 "node" will have 32 CPUs. Consequently jobs that run in 4 nodes or fewer on Phase 1 will be able to run in a single node on the Phase 3 system
- Apart from changing the number of nodes, tasks per node and possibly "consumable resources", there shouldn't be any need to change LoadLeveler scripts submitted to the Phase 3 system
- The latest compilers and libraries will be installed on the Phase 3 clusters, (not the versions in production on HPCA and HPCB e.g. xlf version 8 not version 7). These will be installed on the Phase 1 clusters, but only for testing purposes, not as the default production versions

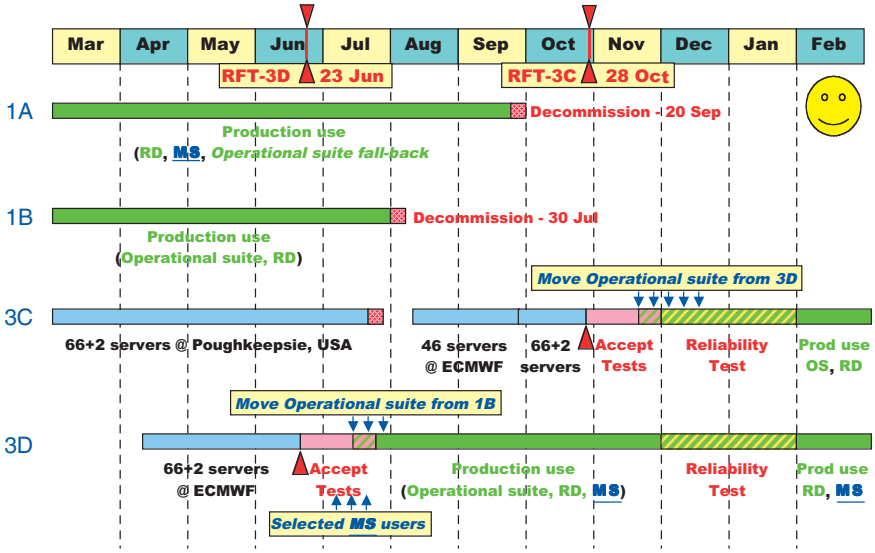
NFS-mounted filesystems

- We have recently been experiencing many problems related to heavy use of NFS-mounted filesystems on the HPC systems
- It was always intended that Filesystems be NFS-mounted on the HPC systems mainly to help with "house-keeping" functions, not for I/O use from batch jobs
- The Operational Suite on the HPC system is being modified to remove all dependencies on NFS filesystems and the Research Department is working towards the same goal for their experiments
- For the Phase 3 clusters we would prefer only to NFS-mount filesystems on the interactive node and would appreciate feedback on this proposal from Member State Representatives

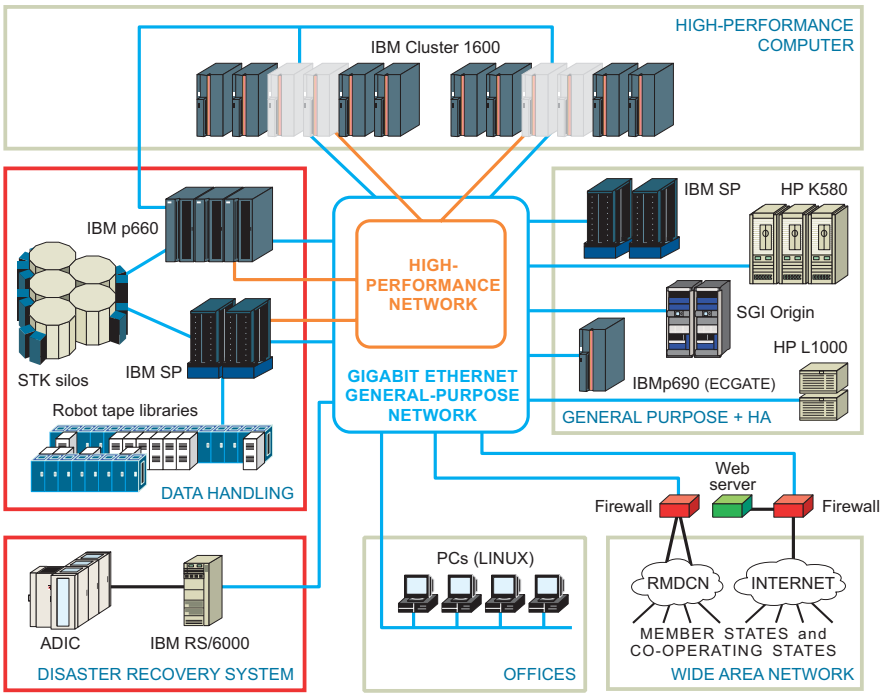
VSD & N/W p690 server (2 per cluster)



Timetable for HPCF Phase 3



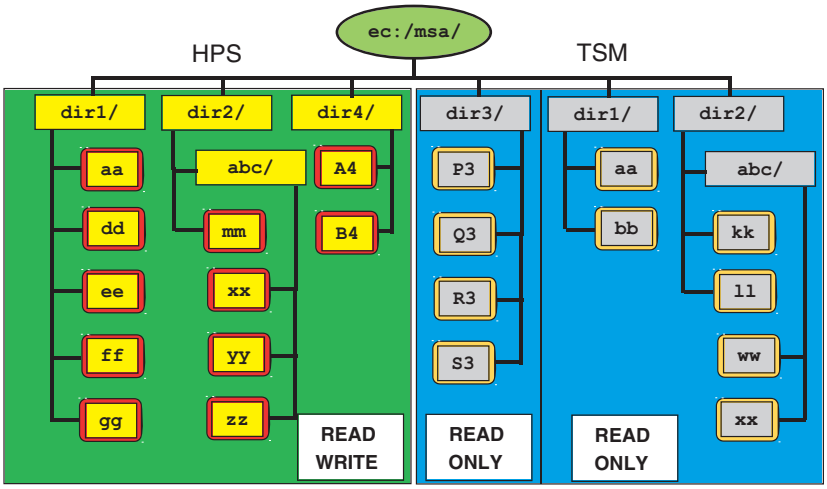
DHS



ECFS migration from TSM to HPSS

- Users who have a lot of data in ECFS have been asked:
  - which of these data can be deleted
  - which of these data they wish to migrate from TSM to HPSS (and of which, of these data, they do NOT wish to have a "backup" copy)
  - which of these data they wish just to keep in TSM until the end of the year (when they will be destroyed)
- With the HPSS version of ECFS no automatic backup copy is made of ECFS data (as there was with the old ECFS system). The user has to specify the "-b" option on the "ecp" command to request a backup copy to be made
- The introduction of the new ECFS service is transparent to users - the new ECFS handles data stored in both the old (TSM-based) and the new (HPSS-based) DHS

ECFS directories in HPSS and TSM



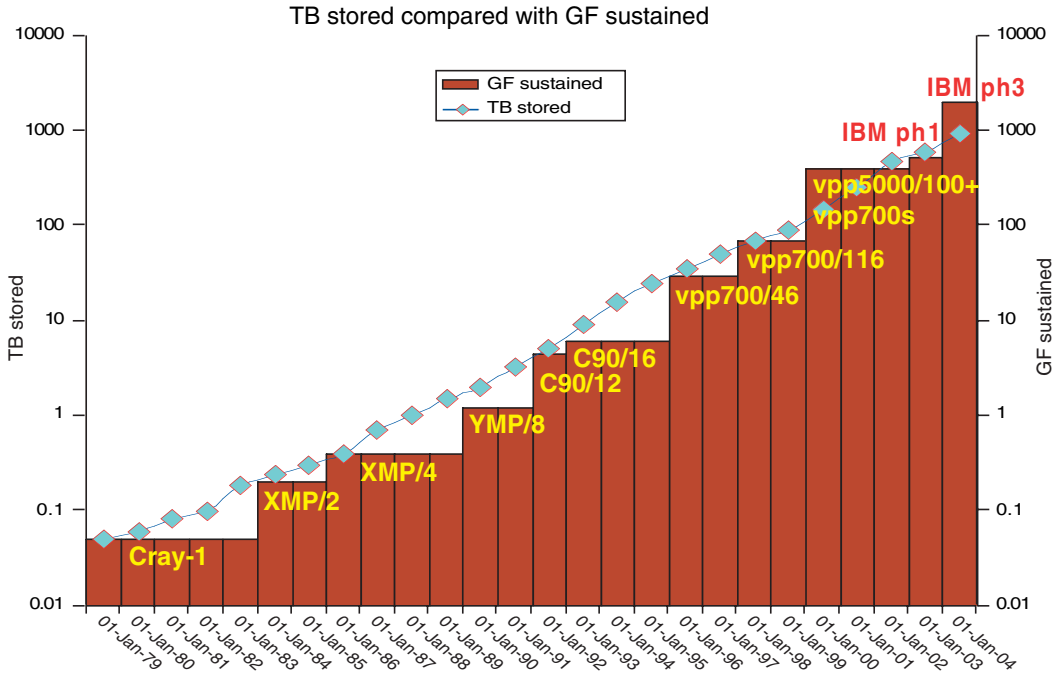
Disaster recovery system - DRS

- ECMWF's data archive is its foremost asset;
- There are many risks associated with data management. Data needs to be protected from:
  - accidental and malicious deletion or damage;
  - bugs in the data management software (vendor's & customer's);
  - disk and tape drive failures and media faults;
  - corruption by S/W and H/W;
  - catastrophes - such as machine room fires, explosions etc;
- Access to the data needs to be resilient to these risks and should also take into account failures of the data servers, storage area networks, robotic tape libraries and such like.

DRS improvements

- ECMWF is currently investigating various options that could improve the capability and functions of the DRS;
- The use of "remote mirroring" of disk volumes to other disks that would reside in the DRS building;
- The use of "flash copy" snapshots to provide point-in-time versions of meta-data;
- Installing sufficient DHS equipment in the DRS building to be able to quickly offer a (degraded) service, should the equipment in the main computer hall be destroyed;
- Upgrading the ADIC AML/J robotic tape library to use Generation-2 LTO drives and media.

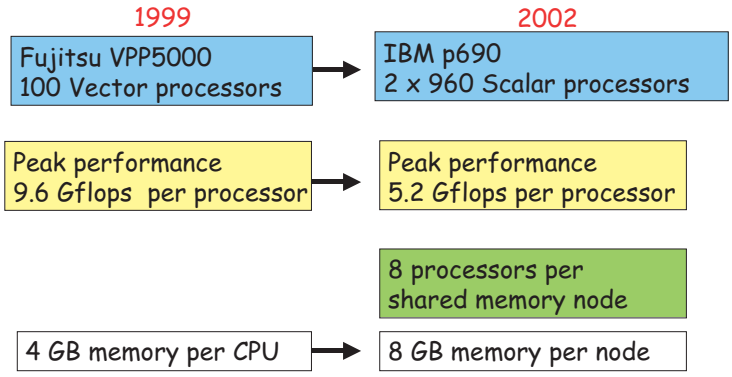
Archive volume compared with computing capacity



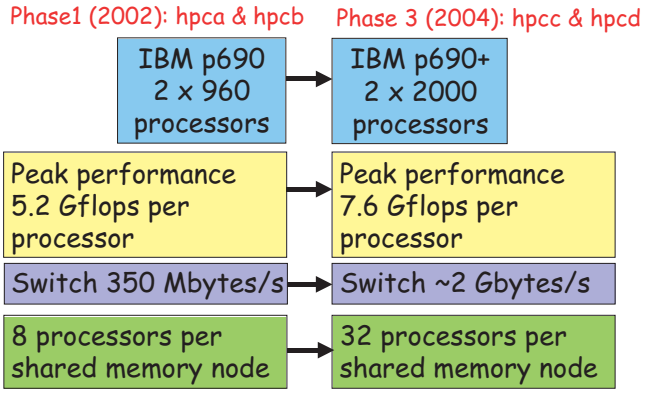
M. Pithon (France) asked, when partitioning stops, with phase 3 of the IBMs, and jobs share nodes, what of contention for other resources, such as memory and access to the Interconnect? N. Storer replied that Workload manager will be used to control access by the jobs to the resources. It can physically limit memory. It allows over-subscription, however, once the oversubscribed resources are required by other tasks, the job using these oversubscribed resources will start to page. This has been reported as a source of performance problems to IBM.

Early experience with Phase 3 test system - *Deborah Salmond & Sami Saarinen*

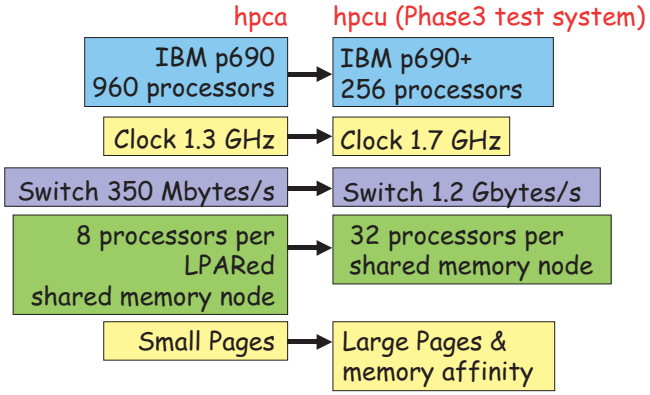
Phase 1 - Migration VPP to IBM



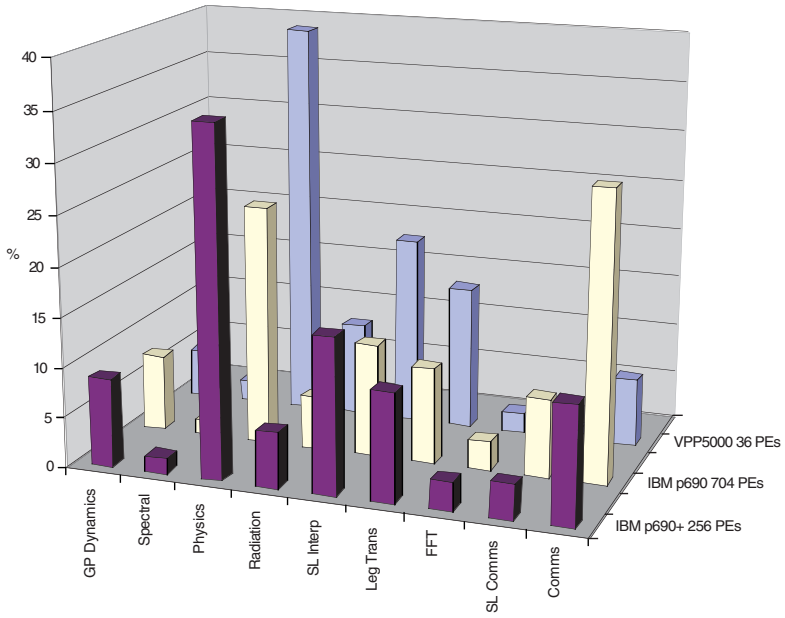
Phase 3 - 4x Performance increase



hpca compared with hpcu



T799 / L90



What is Dr.Hook ?

- A Fortran & C-callable instrumentation library to
  - Trap run-time problems
  - Gather profile info per subroutine
    - Wall-clock or CPU-times
    - Mflop/s & MIPS -rates
- The basic feature: keep track of the calling tree
  - For every MPI-task and OpenMP-thread
  - Upon error (when caught via Unix-signals) tries to print the current active calling tree
  - System's own traceback can also be printed
- Portable with low overhead (~1%)

Dr.Hook environment variables

- Enable Dr.Hook (call-tree/traceback only => cheap)
  - DR_HOOK=1
- Enable wall-clock time profiling information at exit
  - DR_HOOK_OPT=prof
  - The profile will be written to files drhook.prof.<1..nproc>
- Redirect profile-file to /path/file.<1..nproc>
  - DR_HOOK_PROFILE=/path/file
- Restrict output to MPL-task MYPROC=1
  - DR_HOOK_PROFILE_PROC=1
- Collect HPM (Mflop/s & MIPS) information
  - DR_HOOK_OPT=hpmprof or mflops

How to instrument a Fortran90 program with Dr.Hook?

```
  SUBROUTINE SUB
   USE YOMHOOK, ONLY : LHOOK, DR_HOOK
   IMPLICIT NONE

    REAL(8) ZHOOK_HANDLE ! Must be a local (stack) variable

!- The very first statement in the subroutine
    IF (LHOOK) CALL DR_HOOK('SUB',0,ZHOOK_HANDLE)

!--- Body of the routine goes here ---

!- Just before RETURNing from the subroutine
    IF (LHOOK) CALL DR_HOOK('SUB',1,ZHOOK_HANDLE)

    END SUBROUTINE SUB
```

Dr. Hook Traceback

```
 0:  15:57:40 STEP  936 H= 234:00 +CPU= 41.379
 13:[myproc#14,tid#4,pid#55924]: Received signal#24 (SIGXCPU) ; Memory: 2019178K (heap), 0K (stack)
 13:[myproc#14,tid#1,pid#55924]:  MASTER ,#1,st=1,wall=0.000s/0.000s
 13:[myproc#14,tid#1,pid#55924]:   CNT0 ,#1,st=1,wall=0.000s/0.000s
 13:[myproc#14,tid#1,pid#55924]:    CNT1 ,#1,st=1,wall=0.000s/0.000s
 13:[myproc#14,tid#1,pid#55924]:     CNT2 ,#1,st=1,wall=0.000s/0.000s
 13:[myproc#14,tid#1,pid#55924]:      CNT3 ,#1,st=1,wall=0.000s/0.000s
 13:[myproc#14,tid#1,pid#55924]:       CNT4 ,#1,st=1,wall=0.000s/0.000s
 13:[myproc#14,tid#1,pid#55924]:        STEPO ,#978,st=1,wall=10531.259s/0.000s
 13:[myproc#14,tid#1,pid#55924]:         SCAN2H ,#1018,st=1,wall=8913.967s/0.043s
 13:[myproc#14,tid#1,pid#55924]:          SCAN2MDM ,#1018,st=1,wall=8913.896s/32.036s
 13:[myproc#14,tid#1,pid#55924]:           GP_MODEL ,#938,st=1,wall=8845.641s/4.830s
 13:[myproc#14,tid#1,pid#55924]:            EC_PHYS ,#213893,st=1,wall=6144.597s/22.378s
 13:[myproc#14,tid#1,pid#55924]:             CALLPAR ,#213893,st=1,wall=5856.788s/88.130s
 13:[myproc#14,tid#1,pid#55924]:              SLTEND ,#213893,st=1,wall=662.390s/179.559s
 13:[myproc#14,tid#1,pid#55924]:               CUADJTQ ,#117188599,st=1,wall=1992.364s/1477.382s
 13:[myproc#14,tid#4,pid#55924]:                EC_PHYS ,#213356,st=1,wall=6145.442s/22.418s
 13:[myproc#14,tid#4,pid#55924]:                 CALLPAR ,#213356,st=1,wall=5860.376s/88.000s
 13:[myproc#14,tid#4,pid#55924]:                  CUCALLN ,#213810,st=1,wall=2731.710s/27.983s
 13:[myproc#14,tid#4,pid#55924]:                   CUMASTRN ,#213810,st=1,wall=2679.495s/36.678s
 13:[myproc#14,tid#4,pid#55924]:                    CUDDRAFN ,#213810,st=1,wall=66.548s/23.442s
 13:
 13:  Signal received: SIGXCPU - CPU time limit exceeded
 13:
 13:  Traceback:
 13:    Location 0x0000377c
 13:    Offset 0x0000009c in procedure _event_sleep
 13:    Offset 0x00000318 in procedure sigwait
 13:    Offset 0x000006c8 in procedure pm_async_thread
 13:    Offset 0x000000a4 in procedure _pthread_body
 13:    --- End of call chain ---
```

hpcu compared to hpca for T511 L60 forecast run on 128 PEs (32 MPI tasks x 4 OpenMP Threads) at Cycle 28r1

Environment Variables for hpcu

@ network.MPI=css0,,us ‡ for hpcu -> John Hague IBM
@ network.MPI=csss,,us ‡ on hpca

#—- for Memory Affinity ——
export MEMORY_AFFINITY=MCM
export MP_AFFINITY=MCM

#—- for comms performance —
export MP_EAGER_LIMIT=64k
export MP_USE_BULK_XFER=yes
export MP_BULK_MIN_MSG_SIZE=50000

#—- for MPI + multiple OpenMP threads —
export MP_WAIT_MODE=poll
export XLSMPOPTS=”parthds=$omp:stack=$stk : spins=1 : yields=1”

#—- for MPI + 1 OpenMP thread —
export MP_WAIT_MODE=sleep
export XLSMPOPTS=”parthds=$omp:stack=$stk : spins=500000 : yields=50000”

Dr. Hook for T511 forecast - hpca

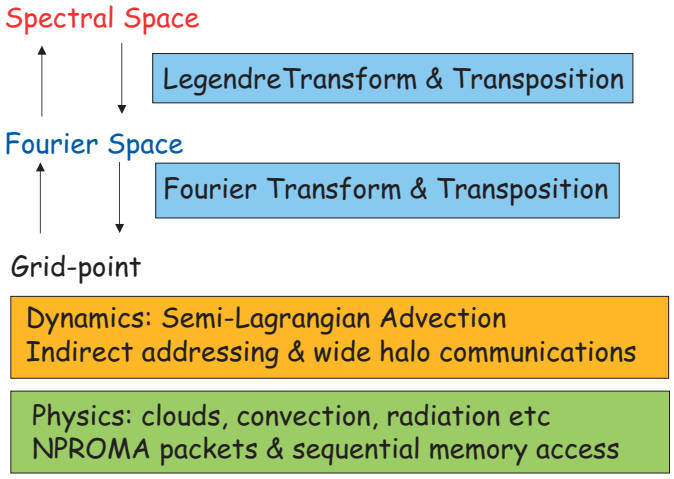
#	% Time (self)	Cumul (sec)	Self (sec)	Total (sec)	# of calls	MIPS	MFlops	Div-%	Routine@<tid> [Cluster:(id,size)]
1	7.43	35.027	35.027	40.573	49	961	273	2.9	WVCOUPLE@1 [567,1]
2	3.67	52.349	17.322	17.367	5824	1113	546	3.6	*CLOUDSC@1 [5,4]
3	3.65	52.349	17.204	17.287	5791	1116	548	3.6	CLOUDSC@4 [5,4]
4	3.64	52.349	17.181	17.289	5769	1118	549	3.6	CLOUDSC@2 [5,4]
5	3.63	52.349	17.138	17.202	5770	1117	549	3.6	CLOUDSC@3 [5,4]
6	3.51	68.918	16.569	16.584	54	783	0	27.6	TRMTOL_COMMS@1 [525,1]
7	2.76	81.935	13.017	18.260	51	926	1	2.8	TRGTOL@1 [520,1]
8	2.51	93.763	11.829	11.831	54	742	0	24.8	TRLTOG_COMMS@1 [523,1]
9	2.41	105.145	11.382	30.536	11540	1106	88	3.4	*CUASCN@3 [30,4]
10	2.40	105.145	11.336	30.436	11538	1112	88	3.4	CUASCN@2 [30,4]
11	2.39	105.145	11.274	30.394	11582	1110	88	3.4	CUASCN@4 [30,4]
12	2.39	105.145	11.267	30.072	11648	1113	86	3.4	CUASCN@1 [30,4]
13	2.36	116.296	11.150	11.185	3492	2135	2172	0.0	*MXMAOP@1 [166,4]
14	2.31	116.296	10.897	10.940	3502	2218	2259	0.0	MXMAOP@2 [166,4]
15	2.30	116.296	10.832	10.920	3474	2216	2258	0.0	MXMAOP@4 [166,4]
16	2.29	116.296	10.816	10.910	3484	2224	2266	0.0	MXMAOP@3 [166,4]
17	1.94	125.448	9.152	9.327	27785	1433	682	0.0	*LAITQM@3 [138,4]
18	1.94	125.448	9.130	9.263	27980	1434	679	0.0	LAITQM@1 [138,4]
19	1.92	125.448	9.073	9.256	27715	1432	682	0.0	LAITQM@4 [138,4]
20	1.92	125.448	9.045	9.220	27750	1440	686	0.0	LAITQM@2 [138,4]
21	1.85	134.173	8.725	8.785	5563	985	592	2.2	*SLTEND@4 [297,4]
22	1.85	134.173	8.724	8.777	5596	987	593	2.2	SLTEND@1 [297,4]
23	1.83	134.173	8.654	8.741	5541	986	593	2.2	SLTEND@2 [297,4]
24	1.83	134.173	8.621	8.658	5546	989	595	2.2	SLTEND@3 [297,4]
25	1.82	142.737	8.565	8.580	51	782	0	21.6	TRLTOM_COMMS@1 [524,1]
26	1.80	151.219	8.482	69.102	13	581	22	10.6	RADINTG@1 [207,1]

% Time (self)	Self (sec)	Total (sec)	# of calls	MFlops	Div-%	Routine@<tid>
3.67	17.322	17.367	5824	546	3.6	*CLOUDSC@1
3.51	16.569	16.584	54	0	27.6	TRMTOL_COMMS@1
2.76	13.017	18.260	51	1	2.8	TRGTOL@1
2.51	11.829	11.831	54	0	24.8	TRLTOG_COMMS@1
2.41	11.382	30.536	11540	88	3.4	*CUASCN@3
2.36	11.150	11.185	3492	2172	0.0	*MXMAOP@1
1.94	9.152	9.327	27785	682	0.0	*LAITQM@3
1.85	8.725	8.785	5563	592	2.2	*SLTEND@4
1.82	8.565	8.580	51	0	21.6	TRLTOM_COMMS@1
1.80	8.482	69.102	13	22	10.6	RADINTG@1

Dr. Hook for T511 forecast - hpcu

% Time (self)	Self (sec)	Total (sec)	# of calls	MFlops	Div-%	Routine
4.77	13.334	13.352	5809	703	3.6	*CLOUDSC@3
2.94	8.206	8.222	3486	2879	0.0	*MXMAOP@1
2.83	7.892	8.425	51	0	0.1	TRGTOL@1
2.43	6.789	6.803	5537	744	2.2	*SLTEND@1
2.30	6.428	6.479	27720	946	0.0	*LAITQM@1
2.12	5.908	5.959	27756	1201	0.6	*LARCHE@1
2.07	5.782	5.906	67662	2381	0.0	*VERINT@1
2.07	5.777	13.502	11558	122	3.5	*CUASCN@4

IFS - Communications and Memory access patterns



T511 forecast - hpcu/hpca CPU ratio for top routines

Routine	hpca/hpcu	Description
CLOUDSC	1.28	Cloud physics
MXMAOP	1.32	Legendre Transform
TRGTOL	1.64	MPI buffer pack/unpack
CUASCN	1.38	Convection
LAITQM	1.40	Semi-Lagrangian Interpolation
LARCHE	1.33	Departure point Calculation
VERINT	1.38	Vertical part of Dynamics

T511 forecast - hpcu/hpca COMMs ratio

Routine	hpca/hpcu	Description
TRMTOL	3.87	Spectral to Fourier
TRLTOM	3.70	Fourier to Spectral
TRLTOG	10.02	Grid-point to Fourier
TRGTOL	10.01	Fourier to Grid-Point
SLCOMM2A	14.66	Semi-Lagrangian

Percentage of total time spent in communications:

- hpca 22%
- hpcu 7.5%

Overall speed-up: (from CPU+COMMS) is 1.57

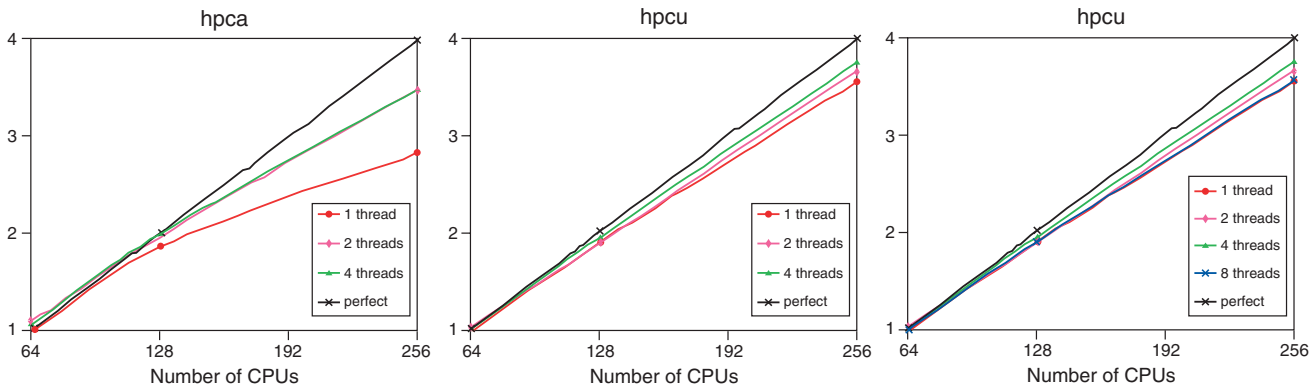
Federation/Colony = 4 & 32 processor nodes

T511 forecast - hpcu/hpca COMMs speeds from mpi_profile - all on switch

Routine	hpca/hpcu	hpcu GB/s (per link)	hpca GB/s (per LPAR)	length MB
TRMTOL	3.87	1.337	0.345	3.79
TRLTOM	3.70	1.268	0.342	1.98

Federation/Colony = 4

T511 forecast scalability with OpenMP threads



Overall performance extrapolations from 128 PE T511 Dr.Hook run

FP operation count for 10 day T511 forecast is 308 Tflops

hpca:

337 Mflops/processor = 6.5% of peak

– 650 Gflops for hpca + hpcb

hpcu:

529 Mflops/processor = 7.8% of peak

– >2Tflops for hpcc + hpcd *

*(assuming each cluster has ~2000 CPUs with 1.7GHz clock & current federation speed)

Arpege : T358 L41 C2.4 (4 day forecast) - Jean-François Estrade

VPP5000	6 CPUs	1740 seconds	
hpca	64 CPUs	16 MPI x 4 OMP	1793
		32 MPI x 2 OMP	1661
		64 MPI x 1 OMP	1784
hpcu	64 CPUs	16 MPI x 4 OMP	1091
		32 MPI x 2 OMP	1036
		64 MPI x 1 OMP	1073

Speed-up hpca/hpcu ~ 1.6 - 1.7

In reply to a question from P. Dando (UK), D. Salmond confirmed that Dr. Hook was available to Member State users. It is useful for any C or Fortran callable code and accepts either, both or neither MPI and OpenMP.

Update on Data and Services - *Baudouin Raoult, Head of Data and Services Section*

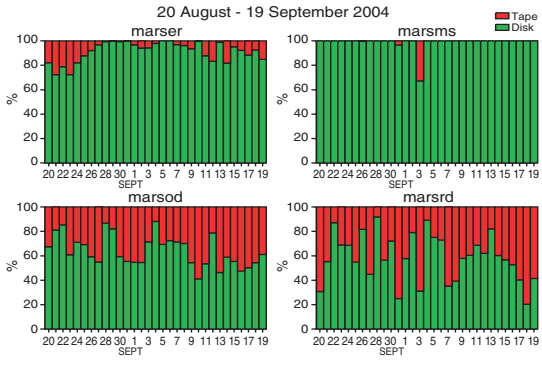
Data And Services Section

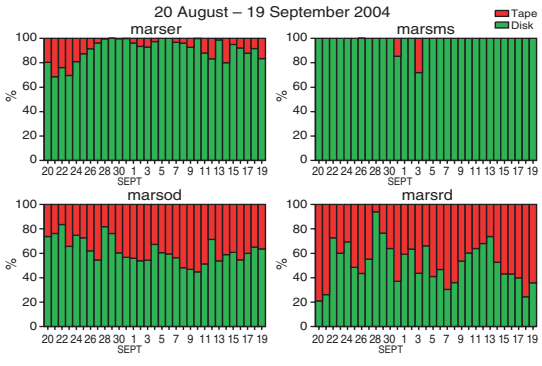
- Ensuring short and long term preservation of the Centre's products
- Providing access to the Centre's data
  - To internal users
  - To the Member States
  - To the research community
  - To the general public
- Managing the Centre's catalogues:
  - Real-time
  - Dissemination
  - Archive
  - Software
- Enforcing data policies
- Providing software to the meteorological community
- Managing licences
  - Data
  - Software
- Support RD, Member States and other research projects
- MARS
- FDB
- GRIB/EMOSLIB
- Dissemination
- Product generation
- Data Services

MARS Update

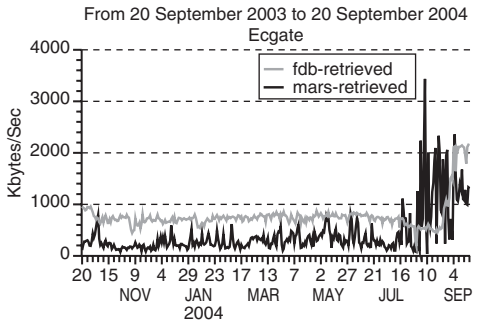
- Migration to HPSS complete
- ERA 40
  - Monthly means, Vertical integrals
  - 4 TB disk space
- New tapes (3592H, 300GB, fast positioning)
- Remote client access using EcAccess
- Implement ERA40 recommended method for wind interpolation

Percentage disk/tape fields retrieved from MARS

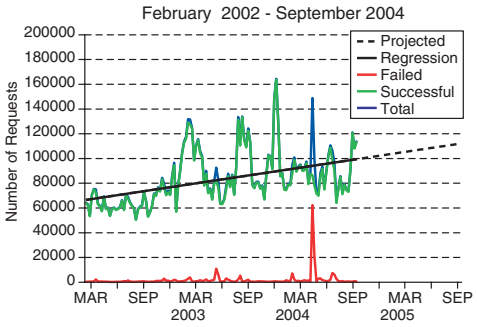




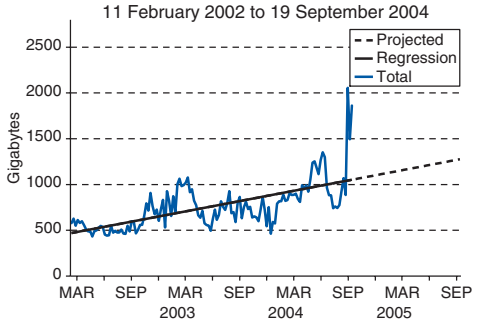
Data retrieved per second



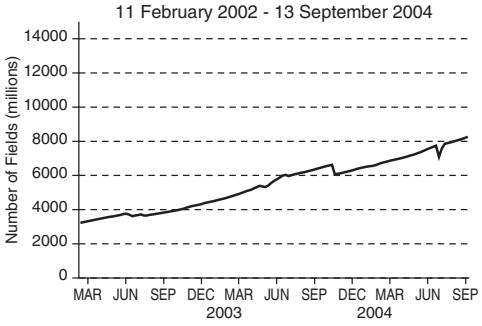
Number of MARS requests



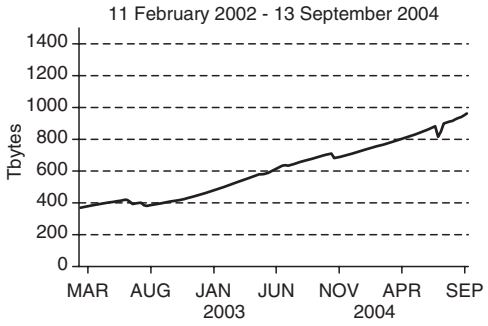
Total MARS data archived daily



MARS archive – total number of fields



Grand total of MARS archive



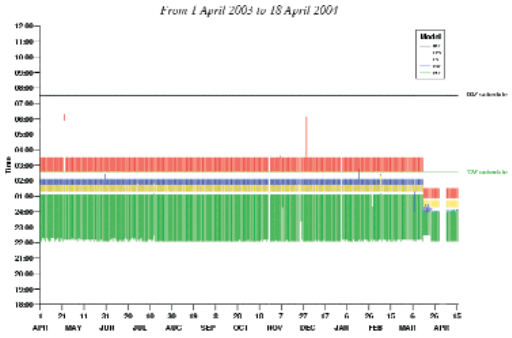
Public Data Server

- Free access for research users
- Linux server with a stand alone MARS server
  - 23 million fields
  - 0.5 Tb of data
- ERA 40 2.5x2.5
  - 2300 registered users
  - 1 Tb delivered monthly
- GRIB to NetCDF
  - In development

Dissemination - New production schedule

- Introduced on the 16th of March 2004
- From 155 minutes to 100 minutes
- Transmission priorities

Daily dissemination times for Germany

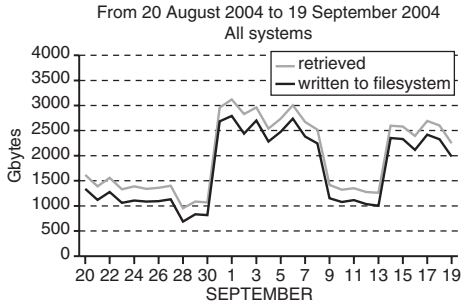


Dissemination - Figures

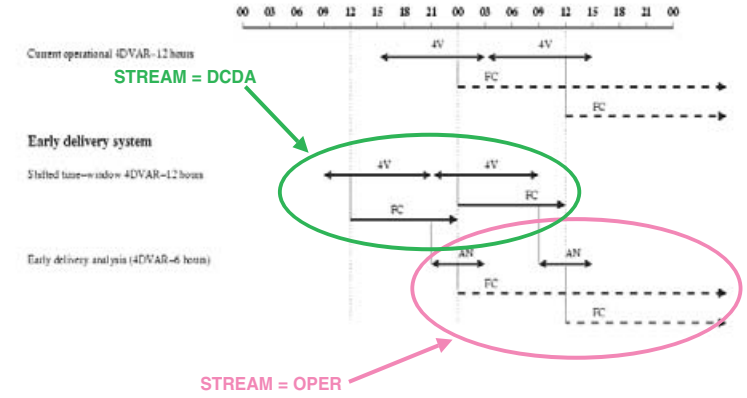
- May 2003
  - 1,200,000 products a day on RMDCN
  - 11.5 Gb a day on RMDCN
  - 85,000 products a day on the Internet
  - 4.4 Gb a day on the Internet
- April 2004
  - 1,650,000 products a day on RMDCN
  - 16 Gb a day on RMDCN
  - 170,000 products a day on the Internet
  - 9 Gb a day on the Internet

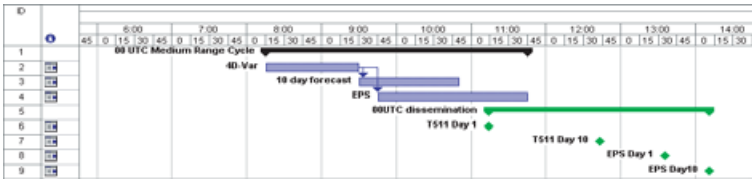
Data Stream	Products (fields)	Files
Main deterministic	460 342	5 789
EPS	927 136	2 096
Wave Global	48 838	1 122
Wave European	4 841	955
Multi Analysis	2 319	665
Short cut-off	389 177	2 326
Wave EPS	27 005	566
Total	**1 859 658**	**13 519**

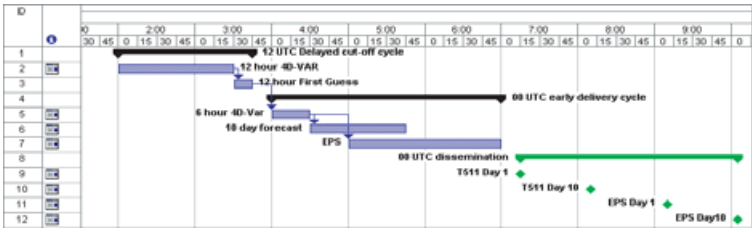
Data accessed from FDB

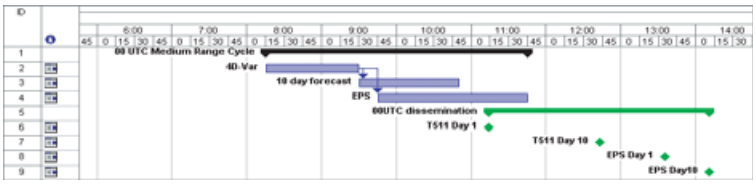


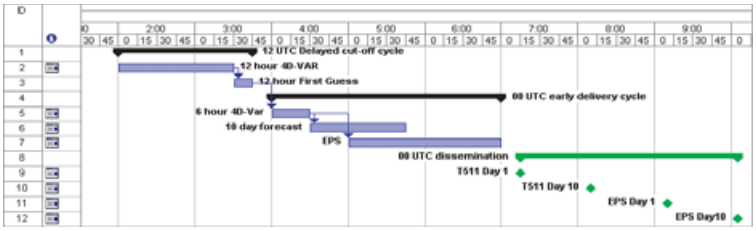
Early delivery system











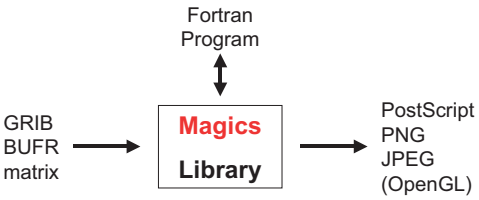
Miscellaneous

- Development of data and software services
  - On-line costing/ordering
  - Packaging of GRIB/BUFR/EMOSLIB
- GRIB2 decoder in development
- New dissemination transport (ECPDS) based on ECaccess in development
  - New monitoring tools

Graphics Update - *Jens Daabeck*

Magics

- Magics is a software system for plotting contours, satellite images, wind fields, observations, symbols, streamlines, isotachs, axes, graphs, text and legends

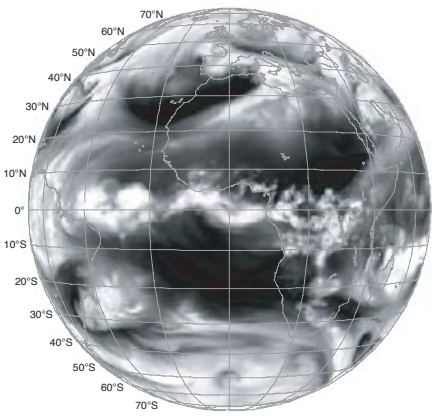


Magics 6.9

New features

- Support for 16-bit simulated satellite images e.g. new 10-bit simulated images
- Reduced dynamic memory allocation handling on all platforms e.g. to support generation of large quantities of maps for Web
- Postscript driver dynamic memory allocation improved
- Correction for the PNG driver (background colour)
- Plans
  - Titles for new data types

New 10-bit simulated image



RTTOV generated radiance SIMULATED Meteosat Image

Magics 6.9 - export

- Available to the Member States
  - 2Q2004
- UNIX platforms
  - Linux SuSE 7.3 (9.1) (Portland Fortran compiler)
  - IBM AIX 5.1
  - SGI IRIX 6.5
  - HP HP-UX B.11
  - HP/Alpha OSF1 V5.1
  - Sun SunOS 5.9
- User Guide in HTML, PDF and PostScript format

The Magics++ Project

- The MAGICS development started in 1984 with the first release in 1985
- To ensure future maintainability, Magics is being migrated to a modern computer language
- Externally, the aim is that existing Magics user programs will need minimal changes to use Magics++
- Phased implementation
- Work has started on the migration of the Magics library from Fortran to C++ including a new contouring algorithm (Akima), implemented in co-operation with INPE/CPTEC

Status

- The internal structure of the new Magics++ has been agreed and the implementation is well underway
- First trials show promising results in using newly developed data decoders (Grib, Grib2, NetCDF) in combination with the newly developed drivers
- A setup has been developed in which the old and new Magics work together to enable a smooth migration between the versions
- To enable easier installation, in the future an automatic script for the configuration, compilation and installation of Magics++ has been developed, based on the widely used autotools
- Limited netCDF support for trial use
- ODB support for trial use
- Plans
  - Better support for Web output with GIF and SVG

The Akima contouring method

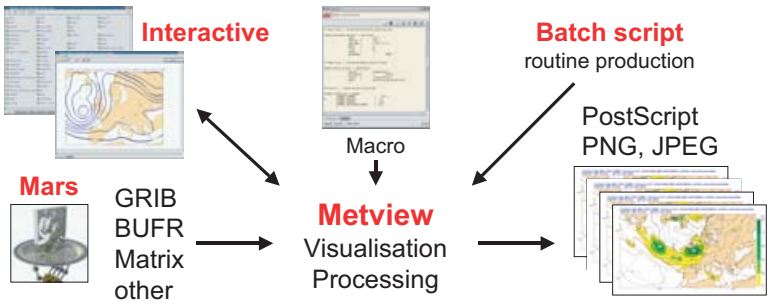
- Three variants of the Akima method were considered for the new contouring package:
  - Algorithms 474 and 760 for generating a denser regular grid, based on an existing grid
  - Algorithm 761 for generating contour lines from an irregularly distributed set of points
- The self-contained versions of these three approaches have been produced in the C++ language
- Algorithm 760 has already been integrated in the Magics++ environment and it is in the evaluation phase
- The other two algorithms are ready to be included in the Magics++ environment

Plan

- Demonstration version of Magics++ was presented at the 9th Meteorological Operational Systems Workshop, 10-14 November, 2003
- Pre-operational release 2Q2004

Metview

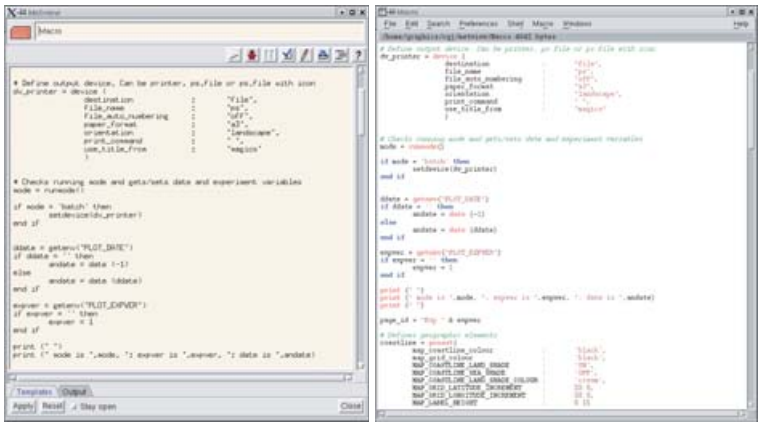
- ECMWF's meteorological data visualisation and processing tool
- Complete working environment for the operational and research meteorologist



New features

- In Metview most work has been done in user support (both for internal and and external Metview users) and in maintenance to implement new user requirements and in fixing bugs
- Work has also been done with a new external Macro Editor which adds several productivity features into Macro editing
- Latest internal Metview version is 3.5.4, based on Magics 6.9, which runs at ECMWF on Linux and AIX platforms

New external Macro Editor: NEDIT



Plans

- Metview 3.6 - export
- High volume satellite data
- Magics++ support

Metview 3.6 - export

- Available to the Member States
  - 3Q2004
- UNIX platforms
  - Linux SuSE 7.3 (9.1) (Portland Fortran compiler)
  - IBM AIX 5.1
  - SGI IRIX 6.5
  - HP HP-UX B.11
  - HP/Alpha OSF1 V5.1
  - Sun SunOS 5.9
- User Guide online
  - PDF and HTML format

EPS Meteograms

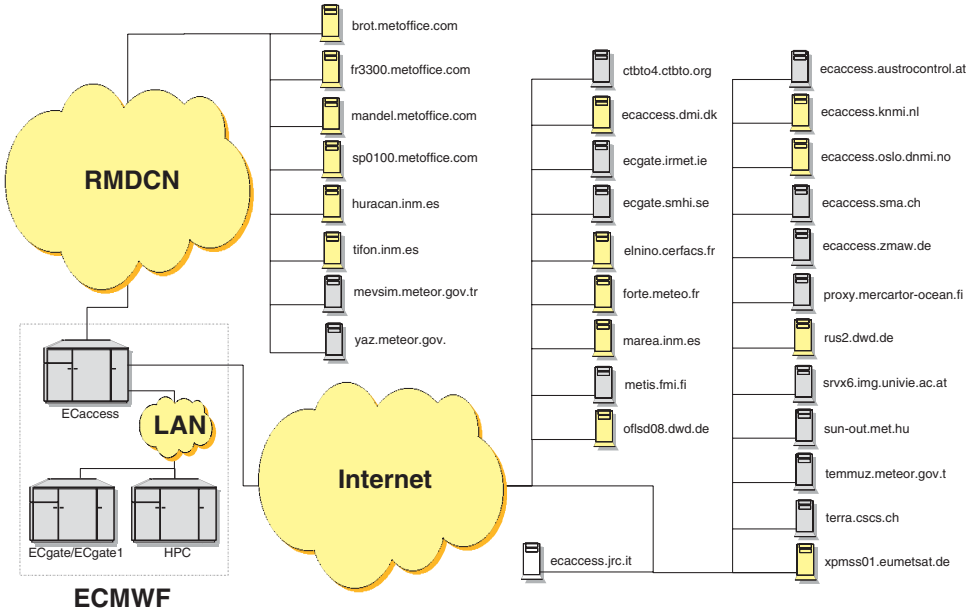
- EPS Meteogram charts available via ECMWF Web pages
  - Shows EPS members forecast distribution for a model run
- Metview user interface
- BUFR data interface
  - Same format as dissemination files
- EPS Meteograms also available as standalone system
- Classic Meteograms available at ECMWF via Metview

Ecaccess, Status & Plans - *Laurent Gougeon*

Ecaccess

- ECaccess provides a portal to access ECMWF archiving and computing facilities
- Strict authentication via SecurID card and X509 certificates
- Data integrity/confidentiality guaranteed by SSL
- ECaccess provides
  - Files and job management in batch or interactive mode through an extended FTP server
  - Files and job management through a Web browser
  - A secure telnet/SSH access to ECMWF
  - A secure X11/VNC access to ECMWF
  - A secure file transfer between ECMWF and systems running the ECaccess Gateway or FTP/SFTP Servers

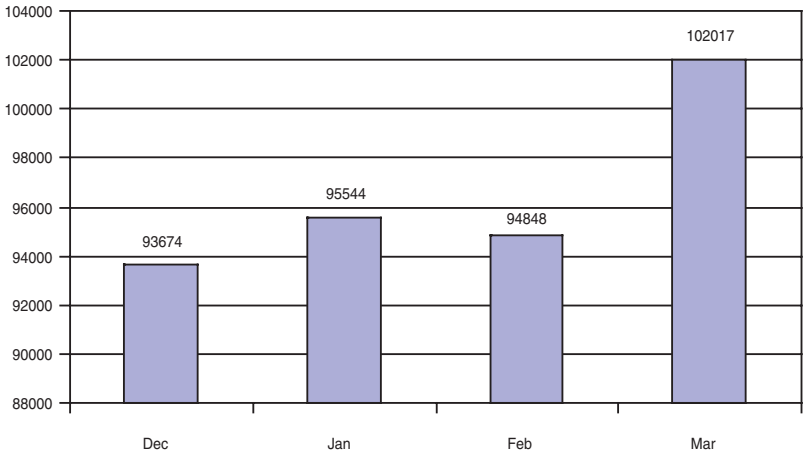
ECaccess Gateways



ECaccess Statistics

Statistics from 04-04-01 to 04-04-15

Gateways	Output	Input	Telnet/SSH
brot.metoffice.com	672.6Mo	-	4 (6m)
ecaccess.austrocontrol.at	-	-	10 (1h)
ctbto4.ctbto.org	6.5Go	74.5Ko	-
ecaccess.dmi.dk	32.1Ko	-	3 (7h)
ecaccess.ecmwf.int	119Go	1.9Go	1745 (4331h)
ecaccess.knmi.nl	37.2Go	502Mo	26 (13h)
ecaccess.oslo.dnmi.no	353.3Mo	-	-
ecaccess.sma.ch	833.2Mo	827.1Ko	-
ecgate.irmet.ie	191.3Mo	-	10 (31h)
ecgate.smhi.se	13.6Go	-	19 (70h)
forte.meteo.fr	53.5Go	916Mo	33 (190h)
fr3300.metoffice.com	19Go	6.2Go	56 (94h)
huracan.inm.es	13.6Mo	-	-
mandel.metoffice.com	705Mo	-	3 (14m)
marea.inm.es	4Go	294.8Ko	26 (124h)
metis.fmi.fi	114.9Go	227Go	-
mevsim.meteor.gov.tr	630.3Mo	-	-
msaccess.ecmwf.int	9.2Mo	-	-
proxy.mercator-ocean.fr	21.7Mo	802.2Ko	9 (21h)
rus2.dwd.de	1.1Ko	-	11 (20h)
sp0100.metoffice.com	1.2Go	8.1Go	7 (4h)
srvx6.img.univie.ac.at	11.6Go	-	-
sun-out.met.hu	704.1Mo	33.3Ko	100 (173h)
tifon.inm.es	13.7Mo	165.3Ko	-
xpmss01.eumetsat.de	1.9Go	-	11 (18h)



Job Submission Enhancements

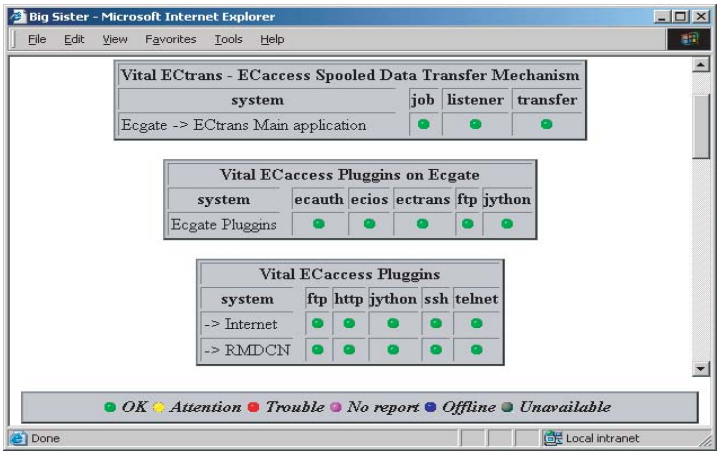
- ECaccess provides Member State users with a common interface to NQS and LoadLeveler
- ECaccess can be used to submit Jobs on different ECMWF platforms
  - The HPC or ECgate (AIX)
  - ECgate1 (IRIX)
- The ECaccess Batch Job Execution system is fault-tolerant
  - Based on the ECtrans spool mechanism
  - The submission machine can be rebooted without losing jobs
- The ECaccess interface allows users to submit scripts with or without scheduler directives to an ECaccess queue
  - A mail notification mechanism is provided

Other Enhancement

- ECaccess symmetry
  - Member State users can use the ECtools at ECMWF using their own shell account as on their workstation
  - Authentication is based on their UNIX account
- Telnet/SSH enhancement
  - From any ECaccess Gateway, users can select the target platform on which they want to login
    - The HPC, ECgate or ECgate1
  - A VNC or X11 proxy can be requested
    - During the login phase with telnet
    - On the command line with SSH
- Access to the MARS archive
  - A release of Metview includes the support of the ECaccess based “ecmars” client

ECaccess Monitoring

- Web interface to monitor ECaccess (Big Sister)
  - Provides the Operators with a simple view of the current ECaccess Network status
    - The ECaccess Servers (Internet and RMDCN)
    - The ECaccess Gateways



  - Notifies when ECaccess is becoming critical
  - Generates a history of status changes

ECaccess Plans

- Dissemination

- The new dissemination system will be combined with ECaccess to allow users to transfer data through the ECaccess Network
- The ECaccess Web Interface will allow MS administrators to monitor their disseminations
- Running the ECaccess and MSaccess Servers and local Gateways in High Availability system will be considered

- ECtrans Enhancement

- Produce a Globus FTP ECtrans module

- Service routing

- Allows the administrator of the Gateway to specify a network (Internet or RMDCN) per service

T. Lorenzen (Denmark) asked whether the service routing mechanism could also allow the specification of Internet or RMDCN depending on the individual Member State user requesting the service: L. Gougeon replied that it could be considered, once the generic service had been established.

User Registration: Update and Demonstration - *Petra Kogel*

Concepts

- The new system: EMS = Entity Management System
- Entities:
  - Users, applications, web domains
- 2 core data sets in a database:
  - User data:
    - Who they work for: ECMWF, specific Met Service, specific university, WMO, ..
    - What they work on: Projects
    - Roles they have: System administrator, Computing Rep, ...
  - Rules: What you do decides the access rights you get. These rules are called “Policies”.

Registration process

- Enter data
  - Bring up web interface
  - Access to web interface strictly controlled
  - Enter user data: name, employer, phone, contact email, ..
  - Tick projects the user works on
  - Choose primary Unix group (if login access is required)
  - Specify additional requests:
    - Login access to ecgate / ecgate1?
    - Login access to HPCA?
    - Access to real time forecast data?
- Press “submit” button
- Request will be checked for validity by central EMS system
- Ok ->
  - A request id will be returned (on screen)
  - The registrations to Unix, web, ECFS, HPC, Mars .. are processed by a batch system
  - Status of request can be seen on-line (started, running, finished)
- Request finishes -> mails are sent
  - To person who performed registration: what has been done
  - To the user who was registered: user name, initial Unix password, web password, SecurID card number, instructions
  - To person dispatching SecurID card
- Immediate request failures
  - Bad request: Failure immediately after “submit”
    - Correct request, eg. Pick different user name if the one chosen is already in use
    - Tell us if you think what you do should work, and that the system is wrong!
  - Permanent failures after initial check: should not exist!
- Temporary failure (system sessions, software bugs):
  - Monitored at ECMWF
  - Fixed at ECMWF
  - Means the registration takes longer than normal (= a few minutes)
  - Does NOT mean that the registration will fail
  - Should be invisible to the Registrator!

Paperwork

- Change current forms to reflect "Rules based system"
- New forms customised for each authorising organisation:
  - National Met Service
  - Special Project Principal Investigator
- Process by which user contacts authorising organisation is unchanged
- User can accept "ECMWF terms & conditions" by logging in to ECMWF web, confirming
  - Acceptance of the terms and conditions
  - Receipt of the SecurID card if applicable

Web Registration or Paper Form?

- Both possible
- Web turnaround should be much faster!
  - Web forms dynamically created
  - Input on first page defines options on following pages

Availability

- Core system went operational in December: used for all registrations since then
- When will there be Member State web registration?
  - As soon as the remaining services (hpc, web, ecfs, SecurID) have been connected to EMS, that is:
    - The registration to these services can be executed in batch mode
    - Use the rules stored in the EMS database
  - ~ Summer (2004)

P. Halton (Ireland) asked whether EMS would also apply to Special Projects. P. Kogel replied that it would: Principal Investigators will have the right to register users for their particular Special Project only.

P. Dando (UK) asked who would fill in the form. P. Kogel replied that a form, tailored to each NMS's requirements would be created dynamically. The user can fill this in and return it to the Computer Representative.

P. Dando asked that Computer Representatives receive an e-mail confirmation of registration.

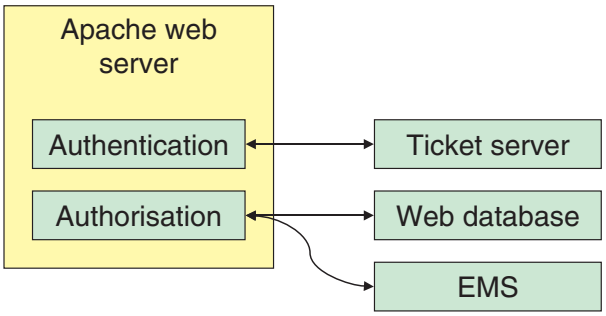
R. Rudsar (Norway) asked whether authority to register users could be delegated to a back-up person. W. Zwieflhofer commented that this was still under discussion, as there were legal aspects to be considered. One possibility might be that ECMWF acts as deputy during a Computing Representative's leave or illness.

Web access control changes - *Carlos Valiente*

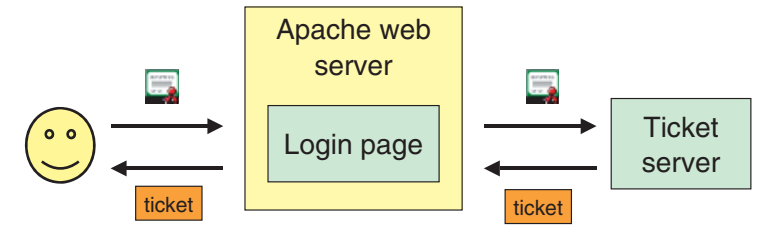
Motivation

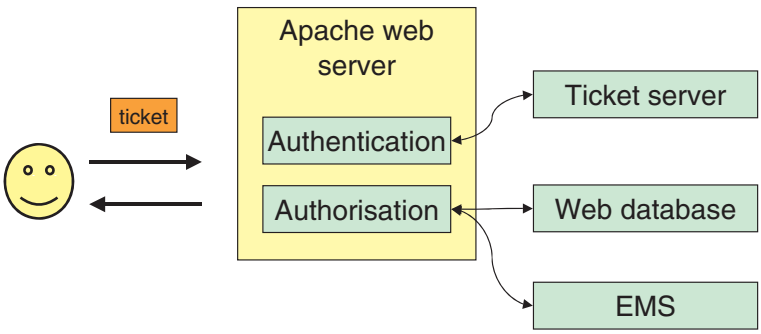
- Improve current system
- Integration with EMS
- Share user identification with dynamic web applications

Overview



How it works

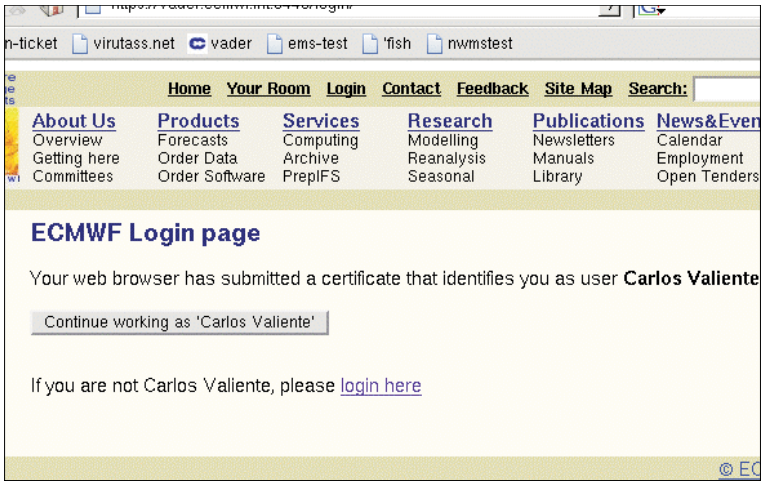




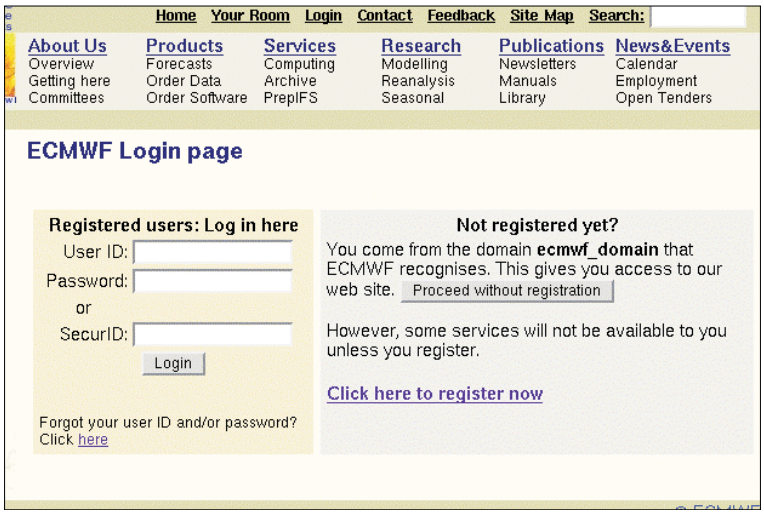
Impact on users

- No transparent login
- Anonymous domain users will have to register in order to access “Your Room” and WebMARS

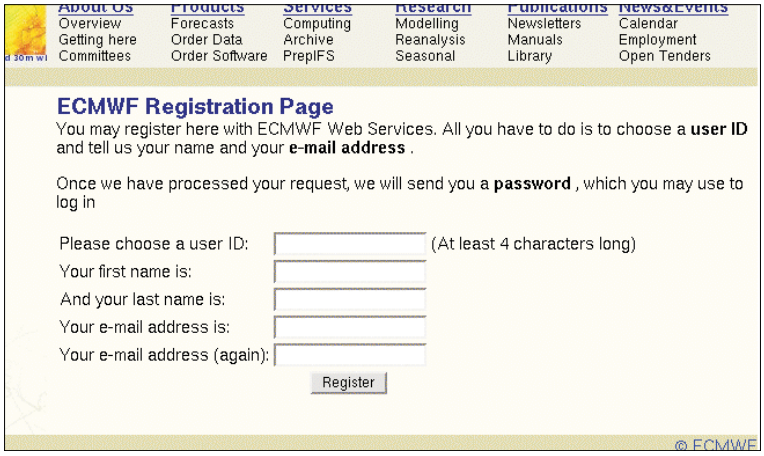
The login page (1)



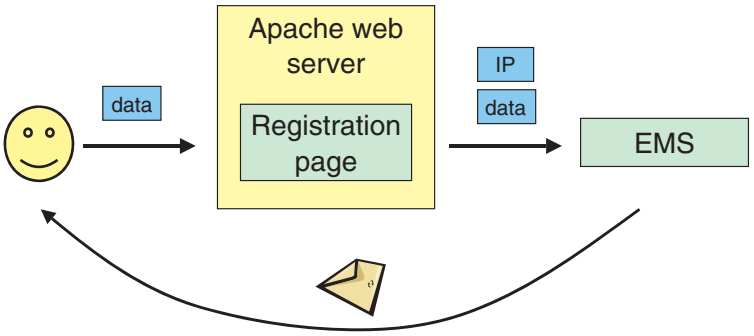
The login page (2)



The registration page



The registration process



What does NOT change

- Access permissions for existing users
- Access permissions for self-registered users: same conditions as domain access (but they DO need to self-register)

W. Zwieflhofer asked Computer Representatives whether they were prepared to allow external registrations, for instance from home, once the original authentication/registration had been carried out in the NMS domain. M. Pithon considered this acceptable.

P. Halton also agreed that it should be possible, but requested that ECMWF keep a log of accesses to monitor potential unauthorised accesses, e.g. to confidential Council documents.

H. de Vries (Netherlands) also supported external access. W. Zwieflhofer said that this would not be possible in the initial version but would be considered as a possible future enhancement.

R. Rudsar asked whether users would leave a trail of tickets. C. Valiente replied that all tickets had a limited validity time and would eventually expire, depending on the application. For instance, web SMS and webprepIFS authorities expire after approx. one hour. A normal user registration via certificate expires after approx. 72 hours and login via user ID and password has almost indefinite validity. R. Rudsar considered that the expiry period for roaming users should be relatively short.

Survey of external users and status of ECgate migration - *Dr Umberto Modigliani*

Aim of the survey

- Determine the level of user satisfaction with the computing services provided by the Centre
- Identify issues of current concern
- Gather quantitative and qualitative data
- Improve the service offered
- Help ECMWF to serve users' needs better

Organisation of the survey

- Send the questionnaire to all registered users who have access to the Centre's computing facilities, i.e. about 1300 users
- The questionnaire will be on the web, to be completed electronically
- There will be pull-down selection lists, checkboxes, etc. for standard answers to questions and "free format" text fields for comments and additional information/suggestions

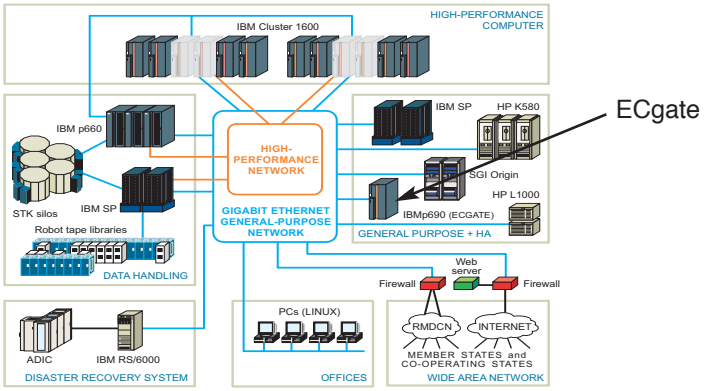
Contents of the survey

- Several sections covering different aspects of the service provided to users
  - information describing each user's activity, work at ECMWF, technical knowledge, etc
  - general evaluation of the services offered
  - use of the ECgate1 server and HPCF
  - more detailed use of the MARS service, including webMARS
  - use of ECFS
  - more detailed use of web services
  - evaluation of User Services
  - area to make suggestions for possible improvements
- There is NO need to answer every question
- The number of questions to answer depends on the specific response to certain questions (branching)
  - the more active/less satisfied users are, the more questions they are asked.
- Users can remain anonymous and information provided will be treated confidentially

Future plans

- Issue the questionnaire as soon as your feedback/comments have been included
- Give about 3 weeks to complete it
- Analyse the results of the questionnaire and produce relevant reports
- Inform users, Computing Representatives

Computing systems configuration



ECgate configuration

- 1 p690 server with **16 CPUs** and **32 GB** of memory
- Each CPU is a Power4 processor running at 1.3 Ghz (5.2 Gigaflops peak)
- The new system is about **3 times** more powerful than ECgate1
- About **1 TB** of usable disk space is provided through a FASTt700 Fibre Channel Disk Subsystem
- File systems use RAID 5 for speed and protection

Migration status

- Documentation and job examples created
- Announced trial access starting on 3 December 2003
  - System was accessible using "**rlogin ecgate**" from ECgate1
- **Full user access** started on **21 January 2004**
  - Direct access using ECaccess/MSaccess available
- Training course organised
  - attended by over **20** Member State users
- Informed **all** registered users **individually**
- Advice/assistance given to several users, in particular those who have been quite active on ECgate1 in the past 3 months
- The system has been quite **stable**
- prepIFS environment being moved to ECgate
  - performance much better
  - solved some issues with a Dutch user
- HIRLAM environment being migrated to run on ECgate
- submission of Member State jobs via SMS being tested
  - required the implementation of a feature not provided by LoadLeveler
  - jobs will NOT automatically be migrated; users will need to change the relevant headers and check their scripts
  - NQS and LoadLeveler jobs could run in parallel for testing purposes
  - job examples will be available

Access to ECgate server

- **Direct** access via ECaccess/MSaccess available:
  - telnet ecaccess.ecmwf.int
  - telnet msaccess.ecmwf.int
    or
  - telnet ecaccess.meteo.ms (your local gateway)
  - Similarly for ftp access
- It is **NOT** possible to access the system via:
  - telnet ecgate.ecmwf.int
  - ftp ecgate.ecmwf.int

Batch environment

LoadLeveler

- LoadLeveler classes û NQS/NQE queues
- No pipe classes with LoadLeveler

- **No command line** options/flags: specify in job header
- No "**waitqueue**" concept: specify output and error files in job header
- **No class for parallel work** has been set up
- CPU-intensive interactive use of the system is discouraged (30 minutes limit)
- The number of classes has been kept to the minimum, but further classes may be added, limits adapted ....

Simple batch job comparison

- NQE/NQS

```
#QSUB -q normal

#QSUB -lt 1000
#QSUB -lT 1000
#QSUB -o /aa/bb/output
#QSUB -eo
#QSUB
:
:
[script]
:
```

- AIX LoadLeveler

```
#@ class = normal

#@ cpu_limit = 1000
#@ job_cpu_limit = 1000
#@ output = out.$(jobid)
#@ error  = out.$(jobid)
#@ queue
:
:
[script]
:
```

Classes

- 3 classes are defined for user work:

Class name	Suitable for	Limits
normal	most batch work: it's the default class	3 hours CPU time Unlimited Wall time 1 GB memory
express	short jobs, access to real-time data	1 hour CPU time 6 hours Wall time 1 GB memory
long	long and/or large jobs	6 hours CPU time Unlimited Wall time 2 GB memory

Compiling environment

- 32-bit or 64-bit addressing mode binaries and libraries
  - two **incompatible** modes.
  - "-q32" or "-q64" options to the compiler
  - 32-bit mode used for ECMWF local libraries.
  - **Default: "-q32"**
- '**underscore**' for external names
  - "-qextname" used for ECMWF local libraries.
  - **Default: "-qextname"**

Software environment

- ECMWF local libraries:
  - ECLIB, accessible through environment variable $ECLIB
    - **Default is 32-bit reals (4-byte REALs)**
  - EMOSLIB, version 240, accessible through environment variable $EMOSLIB
    - **Default is 32-bit reals (4-byte REALs)**
  - NAGLIB, version 20, accessible through environment variable $NAGLIB
    - **Default is 64-bit reals (8-byte REALs)**
  - netCDF, version 3.4 and 3.5
  - HDF version 4.1
- General software packages:
  - MARS, ECFS, MAGICS, METVIEW, TotalView, NCAR Graphics, Midnight Commander, etc.

Information on the Web

- ecgate home page

  www.ecmwf.int/services/computing/ecgate/
- Several job examples available at

  www.ecmwf.int/services/computing/job_examples/ecgate/
- Updated "Introduction for new users"

  www.ecmwf.int/services/computing/help/new_user/intro_ex/
- Computer user training course material available at

  www.ecmwf.int/services/computing/training/material/com_intro.html

  www.ecmwf.int/services/computing/training/material/com_hpcf.html
- Selected IBM manuals available from:

  www.ecmwf.int/publications/manuals/ecgate/

ANY OTHER BUSINESS

J. Greenaway noted that the UK and Ireland had enquired about alternative data transportation, as their telecommunication lines had occasionally been swamped by huge data transfers. W. Zwieflhofer noted that ECMWF now has the capability to write standard LTO tapes (200 GB per tape). M. Fuentes added that a very large data request had recently been supplied on LTO2 tapes.

W. Zwieflhofer commented that the Member States' need to monitor the ECaccess daemon at their end had been noted. The reference in the Administrator's Guide will be reviewed and clarified, if necessary. L. Gougeon also commented that the Big Sister monitoring system at ECMWF could send automatic emails to the Member States' administrators, warning them of daemon problems. This is also a webserver administering the gateway, which could be accessed by Member States operators to monitor the gateway. L. Gougeon noted, however, that, once satisfactorily installed, the gateways are very stable.

W. Zwieflhofer expressed his satisfaction with the Linux cluster presentations. They had been extremely useful. He proposed that this subject should be resumed at the Representatives' next meeting.

NEXT MEETING

It was unanimously agreed that the next meeting should take place in spring 2005.

Part II

Member States', Co-operating States' and ECMWF Linux cluster presentations

FRANCE FRANCE

LINUX at Météo-France - *Marion.Pithon*

- Since 1999 (mail servers)
- Red-Hat for servers. Mandrake for PC.
- Since 2001, some operational production on Linux servers (OKAPI system: production of climatology products). Choice made for performance reasons (compared to HP servers).
- Currently, Linux servers (Red-Hat) for :
  - OKAPI : 5 servers
  - Meteonet 2000 : 7 servers
  - Mail servers : 17 servers
  - Network and security : 5 servers
  - Development servers : 15 servers
  - Telecommunication system : 10 servers

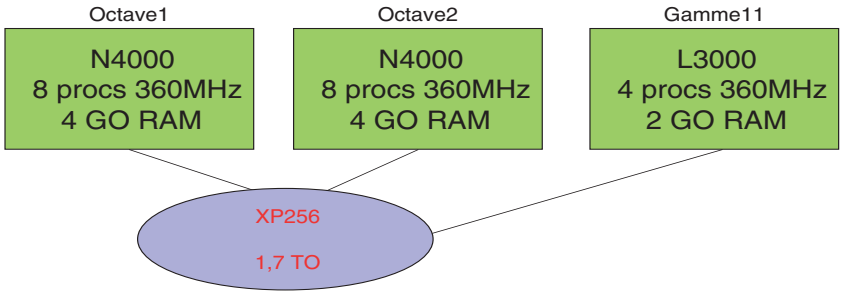
Some issues about O.S

- Change of philosophy (compared to O.S from suppliers)
  - Need of a "professional" distribution? Well supported with a long life cycle.
  - Some software is only certified on "Enterprise" distribution (Oracle on Red Hat Enterprise)
  - Will all distributions be supported and certified by providers?
- Change of organisation (System administrators team)
- What sort of support ? (at M.F minimum of support: contract with a company for 30 calls a year)

Cluster test configuration at Météo-France

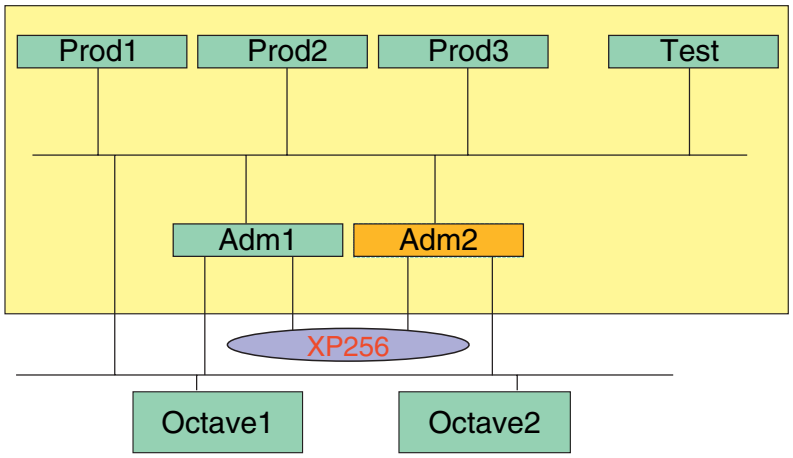
- To replace our current production system
  - Current configuration based on HP servers.
  - Data servers (with database ORACLE).
  - Compute servers (pre-processing).
- Prototype configuration :
  - Cluster of :
    - 4 compute nodes (4 HP DL360 bi pro Xeon2 2GB RAM)
    - 2 nodes for administration and NFS servers (2HP DL320)
    - Switch gigabit for interconnect
    - OS Red Hat 9.0
    - Software ALINKA + OPEN PBS + PGI F90
  - First production ("Immediate forecast" products) in June 2004

Current configuration



FRANCE FRANCE

Prototype configuration



Administration tool : ALINKA

- Administration and management software tool.
- Enables the management of heterogeneous hardware.
- Company : Prologue technology.
- Licence cost is cheap.
- Master/nodes philosophy : one computer, the master, creates and administers many nodes. Nodes are clones of master.
- The cluster can be divided into several logical sub-clusters.
- Web based graphical user interface.
- PBS can be integrated in ALINKA.

General comments

- Need a tool for system management and administration.
- Need a tool for supervision.
- Resources management, scheduling and load balancing.
- Parallel I/O, file systems and storage.
- Middleware for clusters.
- Security.

HUNGARY

HUNGARY

Cluster Project – *László Tölgyesi*

Operative configuration *(test phase)*

PC Linux cluster with 4 nodes (1 master, 3 computing)
Pentium IV(Xeon), 2.4 GHz CPU, 0.5 GB RAM, 40 GB HD per node
GigaBit CISCO switch amoung nodes (1 Gb/s; UTP)
SAN (2 Gb/s), LAN (1 Gb/s) connected to master node
OS: Linux 2.4.20
Cluster SW: OSCAR *(Open Source Cluster Application Resources)*
Loadbalancing, scheduling: Maui *(tested on WEB)*
Job controlling: OpenPBS *(Portable Batch System)*

Note: SGI and HP machines with two Itanium CPU are tested

Planted configuration *(December 2004)*

PC Linux cluster with 17 nodes (1 master, 16 computing)
Two Pentium IV(Xeon), 2.4 GHz CPUs, 2 GB RAM per node

…and later

Linux cluster with 32 or 48 nodes
Itanium 2, 1.5 GHz CPUs, 2 GB RAM per node

IRELAND IRELAND

Linux Cluster Implementation at Met Éireann - *Paul Halton, Head, IT Division*

Why Now?

Main reasons for procuring a Linux Cluster in 2003 were:

- Running costs of the IBM RS/6000 SP [SWIFT] are quite high and the system has had several hardware failures on different nodes.
- More cost effective platform needed for running HIRLAM
- Good in-house experience of using Linux and Open Source S/W
- Success stories from Linux Cluster workshops in Sweden in 2002 (...and again in 2003) & marketplace reports of architecture maturity
- First-hand experience of Linux Clusters running NWP experiments required, before replacing the IBM RS/6000 SP
  - [SWIFT = 9 nodes, each with 4 x 375 Mhz Power3-WH-II CPU's]
- A small budget of Euro50k became available in 2003.

More arguments for buying a Linux Cluster included:

- Backup server needed to run HIRLAM at the same fine-mesh resolution as the operational suite on the IBM RS/6000 SP
- Availability of a cluster would be useful for getting experience of:
  - Managing such a distributed architecture and learning about reliability problems [if any]
  - Porting NWP applications from IBM to Linux Cluster & MPI usage
  - Running new NWP experiments [when the IBM system is fully loaded]
  - Providing in-house computing resources for Special Project, C4I

HPC Marketplace Assessment

Five different machines were identified for comparison:

a) An IBM RS/6000 SP -Power-3 [375Mhz] — similar to our own machine

b) An SGI 3800 R14000 [600Mhz] — similar to a machine used at SMHI

c) A Linux cluster as used in Sweden — [called BRIS]

d) A Linux cluster available from ClusterVision

e) A Linux cluster available from IBM

Relative Speeds

Summary of relative speeds based on information we assembled from various sources...

- SGI 3800 with 16-processors is approx the same speed as a 32-processor IBM RS/6000 SP
- A 16-processor Linux Cluster [BRIS] is about the same speed as a SGI 3800 with 16 processors
- So a 16-processor Linux Cluster is about the same speed as 32-processor IBM RS/6000 SP
- Thus a 16-processor Linux Cluster is about 75% of the power of the 36-processor IBM RS/6000 SP [SWIFT]

IRELAND IRELAND

Relative Costs

Summary of relative costs based on information we assembled from various sources...

- A Linux Cluster would cost about 1/7th the cost of the equivalent
  - SGI 3800 mainframe. (€367,216)
  - IBM mainframe of similar power [SWIFT], (€367,216).
- A Linux Cluster from IBM or ClusterVision, of similar power, would be much cheaper.
- Evidence showed that an MPI version of HIRLAM would run on a Linux Cluster and it would be significantly cheaper. (€53,000 ?)
- Our objective was to start with a small inexpensive cluster to get experience of reliability, scalability, application portability and maintenance & support

Procurement

Procurement of a small 'entry-level' Linux Cluster approved by Met Éireann management in Sept 2003

Invitation To Tender (ITT) issued in Sept 2003 to:

- Dell, IBM, CALYX (Fujitsu-Siemens), SYSNET, ClusterVision, ACT (Sweden) and HP

Proposals were received from:

- IBM, Dell, DSS (HP) and Xpert Technology (HP)

Contract Awarded

- Contract for the supply of a 7-node Linux Cluster was awarded to Dell, in November
- The Dell Cluster was delivered early in December '03
- Cluster is based on the PowerEdge 1750 rack-mounted server

Hardware Specification-1

Dell Linux Cluster (Phase-1) installed Dec 2003:

- 1 x Master node
- 6 x Compute nodes
- 6 x 4 Port Dolphin SCI HBA for Compute Node interconnect
- 1 x Dell PowerConnect Cluster Communication Switch
- 1 x Full size Rack:
- 1 x 42U Rack 4210 Base with a Dell PowerEdge 1750 - AC
- 1 x Backup Tape Unit
- PowerVault Tape System PV112T VS80 Rack Base 1U Single 40/80GB

Hardware Specification-2

1 x Master Node: Dell PowerEdge 1750

- Dual Intel Xeon processors at 2.8GHz with 512kb cache
- 533 Mhz Front Side Bus
- 4GB ECC DDR RAM
- 3 x 146 GB 10k rpm Ultra320 SCSI disk drive
- Dell PERC4/Di U320 Raid controller. (RAID-5)
- On-Board dual PCI-X 10/100/1000 BaseT Ethernet port
- Embedded Remote Access (ERA) port for remote management
- Redundant Power Supply

6 x Compute Node: Dell PowerEdge 1750

- Dual Intel Xeon processors at 3.2 GHz with 1Mb cache
- 533 Mhz Front Side Bus

- 2GB ECC DDR RAM
- 1 x 36 GB 10k rpm Ultra320 SCSI disk drive
- On-Board dual PCI-X 10/100/1000 BaseT Ethernet port
- Embedded Remote Access (ERA) port for remote management
- 4 Port Dolphin SCI HBA for Compute Node interconnect

Software Specification

- RedHat Linux enterprise edition
  - Enterprise Server (ES - Version 3.0) on the Master node and
  - WorkStation (WS - Version 3.0) on the compute nodes
- SCALI Software Suite (Beta-Release installed in Phase-1) included:
  - Scali Interconnect with Dolphin cards
  - Scali MPI Connect which includes:
- Heterogeneous Cluster Support
- Automatic selection of transport mechanism at runtime
- Multithread safe and hot
- UNIX command Line replication
- MIMD support
- Tracing and Monitoring
- Support for Debuggers such as Vampire MPI from Pallas and GNU gdb
  - Scali Manage
- PGI compilers (Fortran and C) & Intel compilers (Fortran and C)

Initial Tests of the Dell Cluster

- After initial installation, tests were run using a temporary Scali licence
- A full release of the Scali software for RedHat was not available until March 2004.
- We only had temporary licences for the PGI Fortran and C compilers.
- All the temporary licences meant that we had to reinstall licences and recompile all applications every two weeks!
- When the PGI compilers were reinstalled, the earlier versions of programs failed to run and error messages similar to the following were produced:
  - EXECUTABLE EXPIRED - This executable was created using a Trial version of PGI software. The PGI Software and derived executables cease to function
- Apart from these inconveniences the system worked very well

Installation Completion (1)

Cluster Installation (Phase Two) March 2004

Scali Engineer and Dell Engineer completed the installation work over 3 days in March 2004

- Hardware:
  - Second Power supply fitted in master node and tested.
  - Tape Drive and SCSI card and cable were fitted.
  - ERA (Embedded Remote Access) installation completed.
  - Each compute node has a 4 Port Dolphin SCI HBA card for compute node interconnect.
  - The compute nodes are connected as a two-dimensional torus via the Dolphin SCI cards.

IRELAND IRELAND

Installation Completion (2)

- Software:
  - RedHat WS 3.0 installed on the slave nodes.
  - Latest version of Scali software installed.
  - Dell OpenManage Version 3.6 was installed and configured.
- Network:
  - Remote system access / management via the ERA/RAC system was configured.
  - Networking software: Scali MPI connect, Scali TCP connect, Scali Manage

Scali Software

ScaliManage™

- Software Distribution & Configuration
- System Administration & Management
- System Monitoring & Alarms

Scali MPI Connect™

- Scali MPI Connect for TCP/IP
- Scali MPI Connect for Direct Ethernet
- Scali MPI Connect for SCI
- Scali MPI Connect for Infiniband
- Scali MPI Connect for Myrinet

Services

- Consulting & Training
- Maintenance & Support

Installation Completion (3)

Benchmark:

- We have been running the full HIRLAM suite on the Dell Linux Cluster continuously for almost 2 months.
- We are very pleased with the performance.
- While it would be interesting to have a LINPACK number, we haven't got the expertise or time to run it at the moment.

Training:

- We require no further hands-on training at this stage!
- We have asked Scali for details of MPI training courses - awaiting reply
- The porting of an identical MPI version of the operational HIRLAM suite was successfully completed during the period Dec 2003 to March 2004.

Installation Completion (4)

Problems encountered during the installation (Phase-2)

- Unable to install latest version of Dell OpenManage - due to incompatibility between Scali Manage & Dell OpenManage.
- The installation of Scali Manage failed - due to a little known bug in the RedHat Package Manager (RPM). Scali logged into our system from Norway and fixed this.
- We need to know how to get around this bug [so that we can upgrade Scali Manage, as newer versions become available]
- Confusion about Netmask addresses. The SCALI engineer, sorted this out in conjunction with colleagues in Norway.

IRELAND IRELAND

Installation Completion (5)
Outstanding issues:

- Unsupported version [i.e. OpenManage version 3.6] of Dell OpenManage presently installed, but it works
  - This didn't hold us up, as we could install slightly older versions of the software
- RPM bug, we need to get a permanent fix
- NFS mounts were lost on one occasion. We are still monitoring this issue.
- Some functionality in the ERA/RAC browser interface does not work, such as:
  - Graceful (smooth) shutdown of system.
  - No response to F2 or ctrl commands - intermittent

Set Up & Management
Benefits of the Scali software

- The RedHat Linux OS was loaded on the Master Node and then we supplied it with relevant details:
  - (IP addresses, node names) and it created the cluster across available nodes.
- The user directories are exported to each of the compute nodes.
- From the desktop, using Scali Manage, we can reboot, power down and monitor the cluster.

Benefits of Dell OpenManage

- Using Dell OpenManage we can monitor the hardware and interact with the BIOS on each node
- To do this on the compute nodes we login to the master node initially
- This is done from a browser interface (we use Netscape)

Cable Management

- Initially when Dell installed the system, cable management arms were installed
- These had to be removed when the Dolphin SCI cards were installed.
- This reduces the ability to slide nodes out & in - the SCI cables have to be disconnected first.

Intel Compilers

- We have an ongoing problem with the licensing software for the Intel compilers.
- The software will not run on RedHat ES 3.0.
- As a work around we are using another Dell workstation (with RedHat 7.2) as the licence server.

Initial Hardware and Software delivery and lessons learned

- Dell appointed a Project Manager from a third Party (SureSkills)
- The nodes arrived from DELL without the ERA ports
- Initially, the Master node only had 1 power supply installed - redundant power supply ordered but only installed in March 2004.
- Dell did not come on site to verify that everything was delivered before installation commenced. This led to delays later.
- Dell learned more about clusters during this installation and a good working relationship was established.
- We purchased the RedHat software directly from RedHat Europe.
- The PGI and Intel Compilers were purchased via Scali. This led to delays, as the PGI software was delivered to Norway and seemed to get lost!
- We recommend that all third Party software should be bought directly in future.

IRELAND IRELAND

Experience (1)

HIRLAM running on the new Linux Cluster

- Currently running version 5.0.1 of the Hirlam forecast model, along with the Hirlam 3DVAR analysis scheme.
- Analysis and model both support MPI.
- Initially a 'stripped-down' version of Hirlam was installed in late December / early January.
- The HIRLAM suite was gradually upgraded to be identical to the operational system running on the IBM RS/6000 SP system

Experience (2)

- The Cluster has been tested using both
  - the Gigabit Ethernet and
  - the Dolphin SCI interconnects.
- The applications were initially compiled using the PGI compilers
- Experiments with Intel compilers have started
- Further optimisation options are under investigation, particularly for 3DVAR analysis

Comparative Timings

Results for the operational Hirlam run on IBM RS/6000 SP were compared over six different runs with results for the same version of the model on the same grid on the Dell Linux Cluster

- The timings are for a 48-hour forecast.
- The grid is 438 x 284 with 31 levels
- The timestep is 300 seconds.

Comparisons made over 6 x HIRLAM runs on each platform:

- The mean time for the IBM SP runs [on 36 CPUs] was 63.3 mins,
- The mean time for the Cluster runs [on 12 CPUs] was 79.5 mins.
- Using these figures gives the result
  - (63.3 / 79.5) = 0.80
  - (79.5 / (63.3) = 1.26
- Thus, the IBM system is 1.26 times as fast as the Cluster or, alternatively, the Cluster is 0.80 times as fast as the IBM system.
- Comparing individual processors
  - (63.3*36) / (79.5*12) = 2.4
  - (79.5*12) / (63.3*36) = 0.4
- Thus a cluster processor is 2.4 times as fast as an IBM processor, or
- An IBM processor is 0.4 times as fast as a Cluster processor.

Comparisons made over 6 x 3DVAR Analysis runs on each platform

- Mean time for the IBM SP runs [on 36 CPUs] was 16.0 mins,
- Mean time for the Cluster runs [on 12 CPUs] was 29.5 mins.
- There is a dramatic difference in performance between the HIRLAM forecast model and the 3DVAR analysis
- The Cluster performance is 80% of the IBM for the HIRLAM model but just 50% for the 3DVAR analysis.
- The 3DVAR we run on SWIFT is extensively optimised, using special IBM libraries for calculating maths functions and for FFT's.
- The Dell Cluster uses a stricter [i.e. slower] implementation of MPI.

IRELAND IRELAND

Future Plans

Currently the full operational NWP cycle includes:

- 3DVAR analysis,
- HIRLAM forecast model,
- Routines for generating climatological files,
- Various post-processing programs,
- A single-processor version of the WAM model and
- Various programs for generating products for customers.

Future work will involve:

- Modifying boundary processing programs to run on more than one node
- Writing a script to check which nodes are available
  - if not all nodes are available, modify the run as appropriate to use the reduced set of nodes
- Implementing the MPI version of WAM.
- Experimenting with the Intel compilers and Maths Libraries to see if they will produce a faster run.
- Adding additional nodes to the Cluster

Summary

- Dell 7-node Linux Cluster with Scali software was installed successfully.
- We have been very impressed with how easy it is to set up the cluster!
- Total cost of 7-node Linux Cluster was €56.6k (ex VAT)
  - H/W Cost: €40.1k
  - S/W Cost: €8.7k
  - Installation Cost: €7.8k
- HIRLAM operational suite (with MPI) was successfully ported to the Dell Linux Cluster
- The entire system has been running reliably for 4 months
- Some minor issues remain to be resolved
- Overall experience to date is very satisfactory.

NORWAY NORWAY

IBM eServer 1350 Linux-cluster at met.no – *Rebecca Rudsar*

Why IBM eServer 1350 with dual AMD Opteron processers and Myrinet?

Three firms (IBM, Dell and HP) came with an offer. They had quite different configurations, therefore we were able to test our benchmark on

- AMD Opteron and Intel Xeon
- Nodes with one and two processers
- Myrinet and Gigabit Ethernet interconnect

In addition we examined the status of 64-bit vs.32-bit processers.

32-bit vs. 64-bit

The code that we run does not need 64-bit memory addressing, as we always run on as many processors as possible and this spreads the memory usage over all the processors. If we needed 64-bit precision in the calculations, this could be done by compiling with 64 bits precision. At present we mange with 32-bits precision and it doesn't seem that 64-bit processors can compete on a price/performance basis yet.

Opteron vs. Xeon

The memory architecture in Opteron appears to be very effective for atmospheric models and the 2.0 GHz Opteron gave a better price/performance picture than a 3.06 GHz Xeon. In addition Opteron has the advantage that applications and Linux can run in 64-bit mode (even though this isn't necessary yet).

One vs. two processers per node

Two processors do not give twice as much performance as one processor but as far as we could see two-processor nodes gave a better price/performance than one-processor nodes even with a Gigabit Ethernet interconnect.

Gigabit vs. Myrinet

The initial start time or latency is very important in ocean- and atmospheric models, since each processor is assigned a physical area (area of the map) and what happens in each area affects to a large degree all the nearest neighbours and to a lesser degree all the other areas. In addition the mathematical algorithms used need global communication. Each single message which is exchanged is, however, not so big and latency is therefore more significant than bandwidth.

Comparing systems with Gigabit Ethernet and Myrinet interconnect showed that over a certain number of processors Myrinet was not only better on performance but also on price/performance. This was because with Gigabit Ethernet the processors were periodically waiting for data from each other.

Why Scali software?

Scali has good administration tools and a good implementation of MPI. Both of these are important but for us the most important thing was the combination of Scali software and OpenPBS queueing system.

We run jobs with different priorities (typically operations, production, research).

All jobs run on the whole cluster. The operational jobs are dependent on using all the nodes to be finished as soon as possible. Thus we had to have a queueing system which makes it possible to suspend jobs with low priority to make sure that jobs with high priority have a fast throughput.

Scali have implemented a solution for this which functions well with the free implementation of PBS, OpenPBS.

Basic software is therefore:

Operating System	RedHat 9 Linux
Administration/Supervision	Scali
MPI & OpenMP	Scali
Compiler	Portland Fortran and C
Queue system	OpenPBS

NORWAY NORWAY

Details of the configuration acquired are given in the main body of Norway's presentation in Part III.

Experience

We have had one occurrance of a broken node. Until then we had been running the jobs on all the processors, i.e. the script which started the job specified the number of nodes to 40. The job expected all the nodes in the correct order and aborted when one node was missing.

To avoid this problem the script was changed to check the nodes which were available and submit the job with a list of available nodes.

They had based their configuration planning on the assumption that the typical MTBF for a server was 70,000 hours, so that a 1,000 server cluster might expect a failure every third day: it is important to build in adequate redundancy and to ensure that the effect of a single node failure is minimised.

Clusters produce a great deal of heat. Their 80-node cluster produces 15 kw and they were obliged to buy a new cooling system.

SERBIA MONTENEGRO SERBIA MONTENEGRO

RHMS of Serbia, BEOWOLF CLUSTER - *Vladimir M. Dimitrijevic*

About 1 year ago we started to use a BEOWULF cluster for running the Eta model with better horizontal resolution of 18 km for 5 day forecasts. At first it was 9 node cluster but very soon we upgraded it to 16 and 20 nodes.

BEOWULF Cluster consists of:

- 20 diskless nodes + one reserve node
- Two servers
- Switch module
- Linux OS
- Message Passing Interface (MPI) - for parallel processing

Cluster 3x3 diskless nodes

3x3 diskless nodes

1 Server node + 1 buck up node-server module

Switch module is based on Cisco switch 2950 and Panduit passive components.

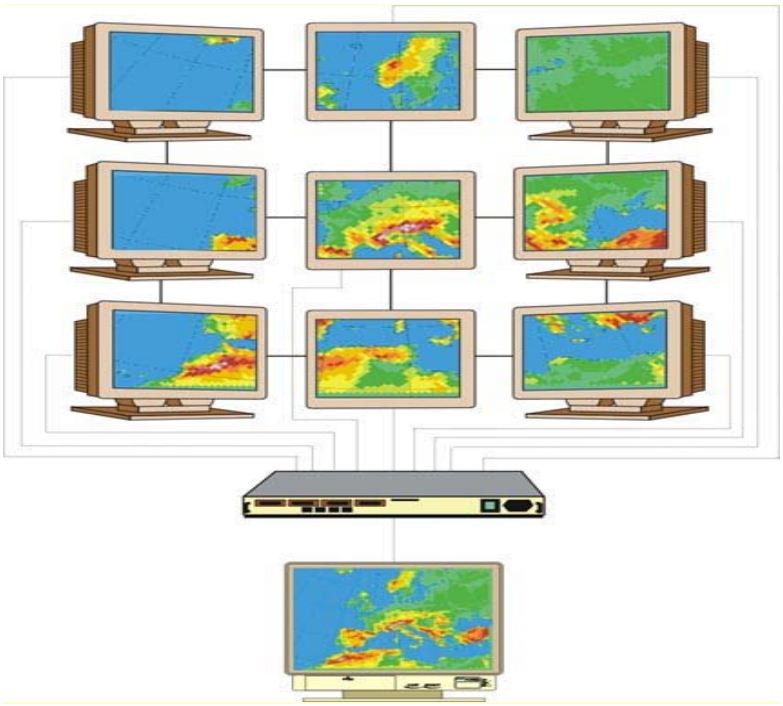
Server and nodes are based on ASUS motherboard and AMD Athlon processor on 1.4GHz

Each node has 512 MB DDR RAM memory and 4 Ethernet ports (one has 5 Ethernet ports) which enables us to link them in simple star or 2-D grid

The operational system is Linux Red Hat 7.2

Nodes share all the resources from the server using the Network File System (NFS).

This configuration works as single computer with 5GB memory, processor performance of 5 Gflops and 100 Mbit Ethernet interconnection between nodes.



ig 1: Here we can see how parallel programming works on a 9-node cluster example. The domain of integration is divided into parts so every node calculates parameters on sub domains and after every time step they exchange the new values of variables. Standard Fortran code of the Eta model is modified using MPI (Message Passing Interface).

Cluster 5x4 diskless nodes

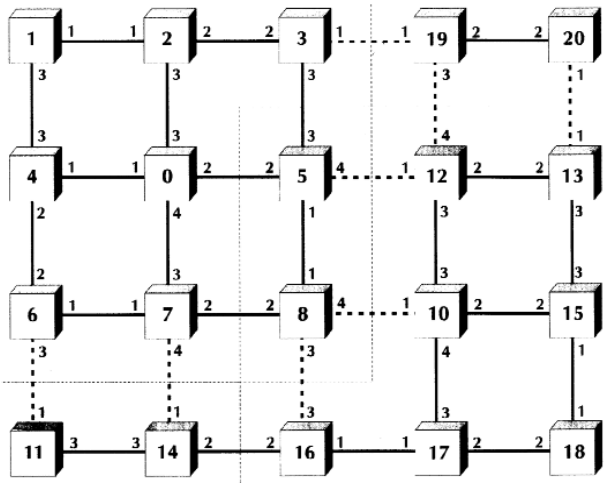


Fig 2: Here we can see how the 2-D grid looks on a 20 node cluster example. The domain of integration is divided in 20 parts, so every node calculates parameters on sub domain and after every time step they exchange new values of variables.

Cluster 2x3x3 diskless nodes

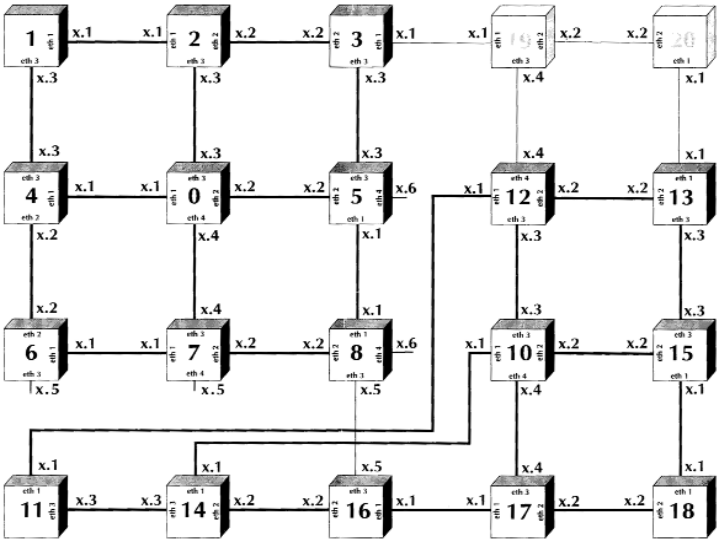


Fig 3: Two separate 9-node clusters can be made easily by switching some cables on the patch panel and disconnecting the two excess nodes. It is also possible to use 4x4 Cluster just by disconnecting 4 nodes (1, 4, 6, 11 or 20, 13, 15, 18).

Experience

- Cluster of 20 nodes showed some irregularities during model integration (infinite values …)
- After testing all the nodes and possible combinations of connections between nodes in 2-D grid, the following conclusion is reached:
  - All nodes work correctly
  - All connections, cables and Ethernet cards are functional
  - Problems with software are excluded (after several runs of different models on same boundary lateral conditions)

Slow Ethernet connections between nodes (100 Mb/s) are a probable cause of problems.

SERBIA MONTENEGRO

SERBIA MONTENEGRO

Conclusion

Cost-benefit of Linux clusters is one of the main reasons for using them.

Considering our needs for running NWP LAM and our financial situation, developing and using Linux Clusters is our best option.

Future plans

- Building the new 6x6 node Beowulf Cluster with Gigabit Myrinet which has 20 times more bandwidth than the 100 Mbit Ethernet that we used.
- Usage of existing 4x4 Cluster or 2x3x3 Clusters as backup.

SLOVENIA SLOVENIA

Linux clusters in Slovenia - *Miha Razinger*

Goals (with NWP in mind)

- "Supercomputing" solution meeting our needs:
  - Enough computing power for our problems
- Taking in account
  - Price of the system
  - Cost of ownership
  - Stability
  - Potential operational requirements
  - Minimized maintenance
  - User friendliness

NWP requirements

- Operational / Research duality
- Computation of forecast in reasonable time on requested domain
- Data assimilation - intensive Input/Output
- Code: Fortran90 programming environment
- MPI / OpenMP programming model
- Specific software (ECMWF - ODB)
- Big amount of data (storage aspects)

History of Linux Clusters at EARS

- 1995 - test of Digital Unix cluster (4 nodes)
- 1995 - demonstrational 20-node Alpha Linux cluster
- 1996 - operational 5-node Alpha Linux cluster
- 2003 - current 14-node Intel Xeon Linux cluster goes operational

Tuba Cluster

- Hardware
    - Processors : Intel Xeon 2.4 GHz (28)
    - 1 master node - SuperMicro 4U Server
    - 13 computational nodes - SuperMicro 2U servers
    - 1 GB memory / processor
  - Storage
    - 350 GB Raid-5 array in Server node
    - 3.5 TB external Raid-5 array
  - Network
    - Gigabit Ethernet - fiber, Entherasys 8000 Gigabit switch

Tuba Cluster Software

- Linux (RedHat 7.3) + Score
- f90 Compilers (Lahey, Intel, Portland Group)
- Totalview debugger
- SMS
- Open source solutions for cluster monitoring (Ganglia)

SLOVENIA SLOVENIA

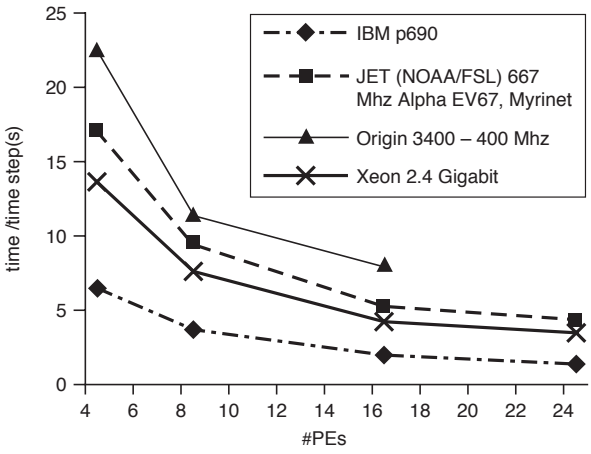
Score (the heart of Tuba)

- Features:
  - LINUX Kernel patch (improved network bandwidth)
  - Modified version of MPICH
  - Synchronized execution of parallel code (gang scheduling)
  - Preempting, checkpointing
  - FIFO scheduler, priority policy
  - Parallel shell
  - Cluster supervision (automatic restart in the case of Score/Hardware failure)

Availability

- Reducing number of single points of failure
- Clusters: great number of redundant boxes
- Usage of Raid-5
- Self correction with Score

Performance



Problems

- Minimal amount of hardware problems
- Some problems due to improper design (master node as a computational node, hyperthreading)
- MPI buffered communications problems
- Score hanging some times

Future Plans

- Challenge: LAM of new generation (2.5 km horizontal resolution), 2007-08
  - at least 10 x times more computing power
- Parallel filesystem (IO) (TerraGrid pfs)
- Faster low latency network (Infiniband)
- Migrate to 64bit (AMD Opteron, Blade servers)
- 64 bit compiler (PathScale)
- Hope to have some results for ECMWF HPC workshop

SLOVENIA SLOVENIA

Conclusions

- Very good price/performance ratio
- With some additional work close to functionality of big systems
- Some unresolved problems with a good hope of resolving them
- Very crucial part (ODB) is still missing for Linux

M. Pithon asked whether the cluster caused much additional work either in its management or from the end-user's point of view. M. Razinger replied that the initial building of the cluster required more effort but that, once running, little maintenance was required and it was very user-friendly. They do not have any external support.

UNITED KINGDOM UNITED KINGDOM

Experiences with Linux Clusters - *Paul Dando, Met Office*

Linux Cluster at BAS

- http://www.antarctica.ac.uk/met/beowulf/
- 13 node (1 master +12 slave) cluster
  - 8 dual processor AMD Athlon MP 1600
  - 5 dual processor AMD Athlon MP 2200
- Ethernet interconnects
- I/O via NFS to 80 GB disk on master node
- No queuing system or job scheduling
- Portland Group Fortran Compiler

Experiences with BAS cluster

- Ethernet interconnect gives poor speed-up beyond 2 processors
- Running Unified Model with 2 processors gives
  - 2 model years/day for global atmosphere-only (32-bit)
  - 0.5 model years/day global coupled atmosphere-ocean (64-bit)
  - Very favourable turnaround compared with other systems
- Only 2 users so do not use batch
- Very reliable
  - 1 PSU failure and 1 memory problem in 2+ years full-time running
- Recent upgrade: 2 dual-node opteron systems + Myrinet

Linux Cluster at RAL

- http://home.badc.rl.ac.uk/iwi/lewis/lewis.html
- 17 node (1 master + 16 slave) cluster
- Each node has
  - Dual processor CPU (2.4GHz Pentium 4 Xeon)
  - 512MB RAM (master has 1GB RAM)
- Slave nodes have Myrinet 2000 networking
- RedHat operating system
- MPICH library for parallel code
- Intel Fortran 90 Compiler version 7
- OpenPBS queueing system
- MAUI job scheduler

Experiences with RAL Cluster

- Some initial problems (PSU, memory failures)
- Ethernet networks used for NFS can be a bottleneck for I/O (better with gigabit ethernet ?)
- Some failures with nodes being unreachable over Myrinet (fixed by replacing the cards)
- Hardware faults on slave nodes do not affect whole cluster (can configure out)
- Overall reliability has been pretty good
- Better performance with fewer processors with more jobs running concurrently

UNITED KINGDOM UNITED KINGDOM

Hadley Centre Linux Cluster

- 26 node (1 master + 25 slave) cluster
- Each node has
  - Dual processor CPU (1.13GHz Pentium III with 512K L2 cache)
  - 1 GB RAM (shared)
  - 40 GB hard disk
- All nodes connected on a private ethernet network using 100baseTX network switch
- Dolphin SCI high-speed interconnect (2-D torus)
- RedHat v7.1, kernel 2.4.3-12scalismp
- Scali Software Platform (SSP v3.0.1)
- Lahey/Fujitsu Fortran 95 Express Release L6.10a

Experiences with Hadley Centre Cluster

- Scaling poor when using both CPUs of a node - otherwise comparable to SX-6
- Running Unified Model (UK Mes forecast)
  - ~1.7 model hours/CPU hour (single CPU/node)
  - ~1.1 model hours/CPU hour (2 CPUs/node)
- Good reliability following initial problems
- Some recent H/W problems
  - Faulty fans, memory, PSUs
  - Recent problems with nodes crashing for large runs

ECMWF **ECMWF**

The ECMWF linux cluster - *Petra Kogel*

Purpose

- Evaluate technologies: Suitable for ECMWF HPCF
- Commission as general purpose server next year

Hardware

- Will be installed in late April
- Supplied by Linux Networx
- Configuration is
  - 32 nodes plus 1 master node
  - Includes 6 I/O nodes with Fibre Channel HBAs
  - Each node has dual 2.2 GHz AMD Opteron 248 CPUs, 4 GB Memory
- InfiniBand low-latency high bandwidth interconnect for MPI

Choices made

- Opteron
  - Performance similar to Power 4+ (as in phase 3 HPC)
  - Xeon: 32 bit only
  - Itanium: Price
- Infiniband
  - Emerging technology, industry standard
  - Price dropping rapidly
  - Excellent expected performance
  - Myrinet, Quadrics: Are known to work, so take the opportunity to evaluate Infiniband now in a "small" setup with limited risk

Shared filesystem

- Which ones?
  - Lustre
  - PVFS 2
  - NFS over fast interconnect
- Problem areas
  - Locking mechanisms / resilience
  - Small and large I/O

Fast Interconnect

- How good is Infiniband?
  - MPI support: Test different versions of MPI
  - Reliability
  - Performance for ECMWF applications

Cluster and Node Management

- Is a production environment distributed over ~4000 feasible?
  - Power down / up: How long does it take?
  - Reboot the cluster:
- How often will that be necessary ?
- How long will it take?
  - Operating system upgrades and patches:
- How difficult?
- Recovery after failure?

Monitoring

- Utilisation, problems, failures, etc.
- Try:
  - Ganglia: Designed for grid, zoom in to cluster and node
  - Other ?

Load balancing

- Batch
- Interactive login: Needed on general purpose server!
- Try:
  - SLURM (default as delivered)
  - Sun Grid Engine
  - Open PBS ???

Other issues

- Support model: Who will fix the bugs?
- Application recovery:
  - The more nodes, the more failures .. Automatic recovery would be “nice”!

PART III

Member States' and Cooperating States' Presentations

AUSTRIA AUSTRIA

Computer equipment

a) Production Server:
SUN Server 420, 2 CPUs/450 MHz, 2GB Memory, Disk 2 GB, CD-ROM
SUN Server 420, 2 CPUs/450 MHz, 2GB Memory, Disk 2 GB, CD-ROM

b) Development Server:
SUN Server 420, 4 CPUs/450 MHz, 2GB Memory, Disk 2*18 GB Raid1, CD-ROM

c) Fileserver:
NET APPLIANCE Network Attached Storage, Disk 500 GB proprietary Raid (~Raid 4)

d) Short-Range_Database Server:
SUN Ultra Enterprise 450 Server, 2 CPUs/300MHz, 2 GB Memory, Disk 4*9.1 GB, CD-ROM, Floppy 3.5"
SUN Ultra Enterprise 450 Server, 2 CPUs/300MHz, 2 GB Memory, Disk 4*9.1 GB, CD-ROM, Floppy 3.5"

e) Long-Range_Database Server:
SUN Enterprise E3500 Server, 4 CPUs/336 MHz, 2GB Memory, Disk 4*9.1 GB, CD-ROM
SUN StorEdge A3500 Disk Array, Disk 2 x 51*9.1 GB
SUN Enterprise E3500 Server, 4 CPUs/336 MHz, 2GB Memory, Disk 4*9.1 GB, CD-ROM
SUN StorEdge A3500 Disk Array, Disk 2 x 51*9.1 GB

f) ECMWF-Server:
SUN Ultra-10, 1 CPU/440 MHz, 524 MB Memory, Disk 2*19 GB, CD-ROM
SUN Ultra-10, 1 CPU/440 MHz, 524 MB Memory, Disk 2*19 GB, CD-ROM

g) GTS-Server:
SUN Ultra-10, 1 CPU/440 MHz, 524 MB Memory, Disk 2*19 GB, CD-ROM
SUN Ultra-10, 1 CPU/440 MHz, 524 MB Memory, Disk 2*19 GB, CD-ROM

h) Internet- and Product Server:
SUN LX50 Server, 2 CPUs/1.4GHz, 1 GB Memory,, Disk 2*72 GB, CD-ROM
SUN LX50 Server, 2 CPUs/1.4GHz, 1 GB Memory,, Disk 2*72 GB, CD-ROM
SUN LX50 Server, 1 CPU/1.4GHz, 512 MB Memory,, Disk 2*36 GB, CD-ROM

i) Intranet-Server:
SUN Ultra-1, 1 CPU, 65 MB Memory, Disk 10.5 GB, CD-ROM

j) Domainname-, Administration- and Operating Server:
SUN Ultra 5_10, 1 CPU, 132 MB Memory, Disk 5.2 GB, CD-ROM
*SUN Ultra-1, 1 CPU, 65 MB Memory, Disk 4.2 GB, CD-ROM

k) Mail-Server:
SUN Netra T1, 1 CPU/500 MHz, 512 MB Memory, Disk 18 GB, CD-ROM
SUN Netra T1, 1 CPU/500 MHz, 512 MB Memory, Disk 18 GB, CD-ROM
SUN Netra st D130, Disk 2*36 GB

l) Backup- / Archive-Server:
SUN Enterprise 250 Server, 2 CPUs, 128 MB Memory, Disk 26.4 GB
Single Equipment with double Access: DLT Cartridge Roboter (3.5 TB, 4 drives)
Single Equipment: Tape 0.5", 9-track, (6250/3000/1600/800 bpi)
Optical Disk Roboter (4 Drives, 144 Slots re-writeable Magneto-Optical- Disk, 650 MB Cartridge)

m) RC-LACE Model Group:
Digital Personal Workstation 600 AU, 1 CPU, 1 GB Memory, Disk 8.6 GB, CD-ROM, Tape 4 mm DAT
SGI Origin 3400, 20 x R14000 CPUs/500MHz, 20 GB Memory, Disk 2*18 GB, 8*73 GB, Tape 4 mm DAT

n) FIREWALL:
XXXXXXX, Confidential

and more than 60 other Servers and Clients depending on special needs at the several Departments and Regional Services of ZAMG, and a flock of nearly 300 PCs, some of them used for routine work, e.g. for forecasters and to supply special Media (Broadcast and Television, Newspapers).

AUSTRIA AUSTRIA

Software

SUN-Systems

Operating System: Solaris (UNIX)
Compiler: Fortran 77, 90, 95, C, ANSI C, C++
Script language: Perl
Graphics: Xelion GKS, MAGICS, PV-Wave, OpenGL
Libraries: IMSL, NAg
Database: SYBASE
GIS: ARC/INFO
Backup SW: Veritas Netbackup
e-mail: Z-mail

LX50

Operating System: Sun Linux

Digital Workstation

Operating System: Digital UNIX
Compiler: Fortran 90, C++
Graphics: NCAR Graphics

SGI-System

Operating System: IRIX64
Compiler: Fortran 77, 90, C, C++
Graphics: NCAR Graphics

Personal Computer

Operating System: Windows NT, Windows 2000, Linux (SuSe, REDHAT),MacOS
Compiler: Fortran, Visual Basic, C
Graphics: Xelion GKS, MAGICS
Applications: MS Office, ABC Flowcharter, ARC/VIEW, CorelDraw,Exchange, Exceed, PageMaker, PhotoShop, SYBASE ODBC,OnNet interdrive
Internet/e-mail: Netscape, Internet Explorer, Outlook / Outlook Express

Operational ECMWF-data in Austria

The operational ECMWF-Data for Austria from the 00Z and 12Z model runs are transmitted to Austria by using :

- ECMWF Dissemination System (most products) : 306 Gribfiles and 2 Bufrfiles are sent to Austria every day (one file per forecast step)
- MSJ-Jobs : EPS-Meteograms as postscript Files, 9 files per day
- special products are downloaded from the internet (EFI-Forecasts)

The ECMWF-Data are sent to the ECMWF-Servers zaaecm1 and zaaecm2 in Vienna

(zaaecm2 is used when the zaaecm1 has broken down).

The daily operational ECMWF-data comprises 330 megabyte. The data is stored ten days on zaaecm[12], a part of it is archived on magnetic tapes.

On the zaaecm[12] the data is checked and copied to the multi-user server zaamus1p (coupled with zaamus2p) for public use as soon as it becomes available and has passed the validation.

AUSTRIA AUSTRIA

Overview : ECMWF-Datastreams used by the Austrian Weather Services:

1 Gribdata from T511 modell, area Europe/North Atlantic/North Africa

Two data streams "A0D" and "A4D" are used (A0D/A4D-Gribfiles from the 12Z and 00Z run) with the following contents :

- 1.5 degrees latitude-longitude grid in the area from 90W-90E and 90N-18N
- analyses 00 06 12 18 UTC and forecast steps 6-240 hours in 6-hourly intervals
- parameters on pressure levels 1000-200 HPA (u/v-wind, vertical velocity, temperature, geopotentiel, relative humidity)
- surface parameters (e.g. temperature/dew point 2M, 10M u/v wind, wind gusts, cloud cover (total, low, medium, high), precipitation (total, convective), mean sea level pressure, snowfall and other)
- additional data from a global 1.5 degrees grid are also available (only a few parameters and the forecast steps 12-84 hours)
- A0D-Data are also used as input for AUSTROMOS (Model Output Statistics for Austria) forecasting local weather for 112 weather stations in Austria and 150 locations outside of Austria
- 44 A0D and 41 A4D-Files are disseminated to austria every day

2 High-resolution gribdata from T511 modell, area Central Europe

Two data streams "A3D" and "A6D" are used (A3D/A6D-Gribfiles from the 12Z and 00Z run), they have been established in operational weather forecasting during the last year and contain :

- 0.5 degrees latitude-longitude grid in the area from 4E-21E and 54N-43N
- analyses 00 06 12 18 UTC, forecast steps 3-72 hours in 3-hourly and 78-240 hours in 6 hourly intervals
- the same parameters as in data streams A0D/A4D, described in (1).
- 56 A3D and 53 A4D-Files are disseminated to austria daily

3 Gribdata on Model Levels (A1D-Data, 12Z run only)

The A1D-Files are used in the Environmental Department to compute trajectories :

- data on model levels 60-26
- 1.0 degrees grid from 90W-90E and 90N-18N
- analyses 00 06 12 18 UTC and forecast steps 6-84 hours (6hr interval)
- 18 A1D-Files are disseminated to austria daily

4 Ensemble forecasts for Europe (A0E-Data, 12Z run only)

The A0E-gribdata are processed with MAGICS and shown as graphical weather charts and contain :

- cluster means, ensemble means, standard deviation and probability fore- casts
- 1.5 degrees latitude-longitude grid in the area from 90W-90E and 90N-18N
- precipitation, temperature 850 HPA, geopotential 1000/500 HPA
- 7 A0E-Files are disseminated to austria daily

5) Weather Parameters (AYA/AYB-Data in Bufrcode, 12Z run only)

The AYA and AYE Data are deterministic (AYA) and EPS-Forecasts (AYB, Ensembles 1-50 and control forecasts) of temperature for Vienna used by the Vienna Main Heating plant ("Fernwaerme"). These are the only products in Bufr Code used in Austria. 1 AYA and 1 AYB-Bufrfile are sent to austria every day.

6) Forecasts from the European Wave Model (ASM-gribdata, 12Z and 00Z run)

The ASM-gribdata contain forecasts of significant wave height and mean wave direction for the mediterranean area for the forecasts steps 6-78 hours in 6 hourly interval.

The ASM-Data have been made operational in April 2004 and replaced the former A0M-Datastream; 24 ASM-Files are sent to austria per day.

7) Special Forecasts for Italy (A2D/A2E-Data, 12Z run only)

The A2D and A2E-Data contain deterministic (A2D) and EPS Forecasts (A2E, Ensembles 1-50) for a grid point in Northern Italy and is sent to ARPA (Regional Weather Service Venezia) per e-Mail (12 to 240 hours after base time in 12-hr intervals). 21 A2D- and 20 A2E-Files are disseminated to austria per day.

8) Precipitation ensemble forecasts for Austria (APE-Data, 12Z only)

The APE-gribdata are used for precipitation weather forecasts sent to the main electricity power company in Austria:

- control forecast and ensembles 1-50 (total precipitation)
- 1.5 degrees latitude-longitude grid in the area from 9E-18E and 49.5N-45N
- forecast steps from 12 to 120 hrs in 6-hr interval
- 19 APE-Files are disseminated to austria per day.

In addition to the grib- and bufrfiles from the Dissemination System, also 9 postskript files are sent to Austria by MSJ-Jobs :

9) EPS-Meteograms (Postscript Files)

EPS-Meteograms are drawn for Vienna and the capitals of the 8 districts in Austria by a user job (ecgate1) and are transmitted to Austria by FTP, printed out and visualized in the Intranet Server.

In the past the generating of EPS-Meteograms has sometimes failed due to new data structures, missing access rights and installation of new Versions of Metview. It is under discussion to take the EPS-Meteograms directly from the Internet web pages.

10) Extremely Forecast Index

Graphical forecasts of the EFI are downloaded directly from Internet web pages by the synoptic department and stored on the Intranet.

Processing and Usage of ECMWF-Data

The software consists of Korn-Shell Scripts (Solaris Unix-System) for operational control of the data and Fortran77-Programs for deriving special products. The gribdata are processed by using the GRIBEX-Software.

Although the use of Ensemble Forecasts has extended, the main production data is from the T511 model and the 12Z run. In March 2004 the 00Z run has been established in operational weather forecasting in Vienna also.

The T511 data is processed by the new Fortran 77 program "ECMMOD" (=ECMWF Model Output Diagnosis). "ECMMOD" is applied to actual or historic data of the T511 modell, the 00Z and 12Z run and the 1.5 deg grid Europe/North Atlantic or 0.5 deg grid Central Europe:

- ECMMOD reads and decodes the original gribdata from a specified dataset (date of base time, 00 or 12Z run, a defined grid and forecast steps)
- ECMMOD derives additional parameters (BAZI (Baroclinic Zone Indicator), specific Humididy Index, Temperature, Vorticity Advection, Showalter Index, level with 0 deg Celsius temperature, convective cloud cover, windshearing and others)
- ECMMOD has been implemented on the ECMWF-Servers zaaecm1/zaaecm2 and the multiuser public production servers zaamus1p/zaamus2p

The original and derived data is stored on arrays internally and the output of ECMMOD comprises:

- gridpoint values for postprocessing (Ascii-Files)

Graphical products are produced using the gridpoint data (weather charts for Europe and Austria by using MAGICS) as hardcopies and visualized on the Intranet Server.

Mainly used are charts for Europe/North Atlantic for sea level pressure, geopotential 850 and 500 HPA, equivalent relative topography, humidity index, precipitation, temperature and vorticity advection, baroclinic zone indicator, total cloud cover, temperature 2M and 850 HPA.

- QFA-forecasts for 500 different locations in the whole world (QFA's are ECMWF-forecasts interpolated from the gridpoints to the locations, new is the usage of high resolution data in Central Europe)
- selected products (QFA's and also MOS-Data) are stored in a SYBASE data bench at the ZAMG and the regional departments for 7 days
- the data is coded and stored on gribfiles which will be used in the future as the database for meteorological products (one file per time step, the ECMWF-Standardname)
- also special forecasts for customers and private meteorological companies are derived

Corresponding grib data is also available from the DWD (German Weather Service), based on the obervations at 00, 12 and 18 UTC, but only up to 72 hours after the base time and with fewer elements and no Ensemble products.

The Programm DWDMOD is a copy of ECMMOD and is applied to the DWDData.

Additionally also the Output of the ALADIN-Modell is used (fine meshed model, but only for Central Europe and two forecast days). The programs processing the ALADIN-Data is a program written by the model working group, not by the Computer department.

The new Fortran program "PROGVER" has been developed for verification of selected ECMWF-Forecasts based on 0.5 to 1.5 degrees grid compared with MOS and other reference forecasts such as persistence.

Users of ECMWF Products

ECMWF-data is used by the Austria Weather Services internally (operational and scientific purposes) and to derive products for private customers :

1) Operational use by the Austrian public weather services

- ZAMG: Central Institute for Meteorology in Vienna and the 4 Regional Departments in Innsbruck, Salzburg, Klagenfurt and Graz

The computer department is responsible for ordering, obtaining and processing the ECMWF-data (e.g. weather charts), the synoptic department uses the ECMWF-data as a basis for dayly weather forecasting

- MWD: Military Weather Service of the Austrian army and the air force
- ACG: Civil Aviation Weather Service (Austro Control)

It is discussed to combine the three weather services to a single one and establish a private company named "Met Austria".

2) Scientific purposes (special projects) - actual data from dissemination and archived MARS-Data

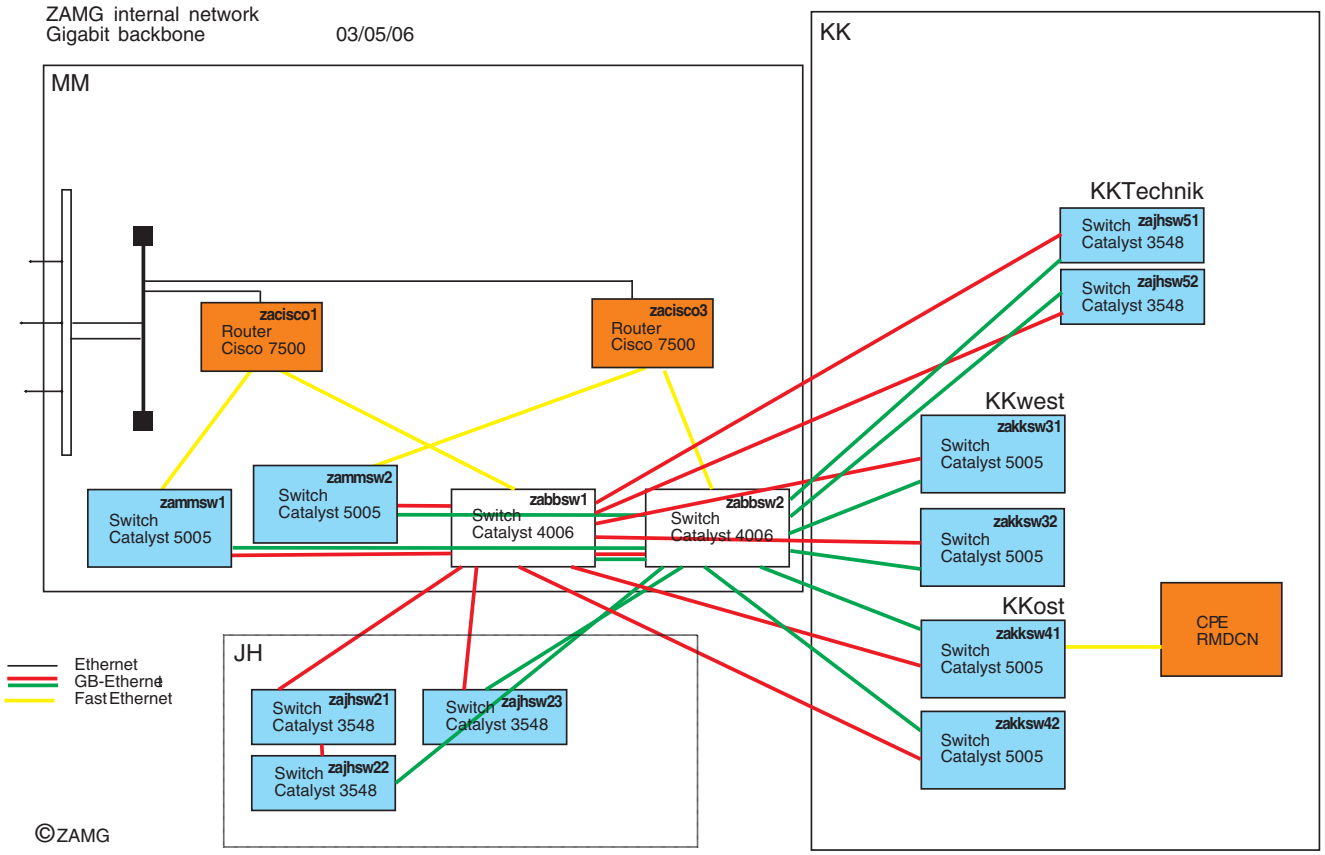
- internal use at the Central Institute of Meteorology in
  - Model Working Group (e.g. for the project "verification")
  - Remote Sensing Group (e.g. combining with satellite data)
  - Environmental Department (e.g. for computing trajectories)
- University Institutes in Vienna, Innsbruck, Graz (e.g. Steinacker, Ehrendorfer, Haimberger, Skoda and advanced students)

3) Private and public customers (only derived products), e.g.

- ORF - Austrian Broadcasting Corporation
- local authorities
- some newspapers
- organizers of sport and culture events
- tourist traffic offices
- street services (esp. snowfall and freezing rain)
- environmental purposes
- electric supply companies (forecasts of precipitation and temperature)
- warning of extreme weather situations (placed in Internet) such as strong wind, extreme precipitation amounts, thunderstorms, icing conditions

AUSTRIA

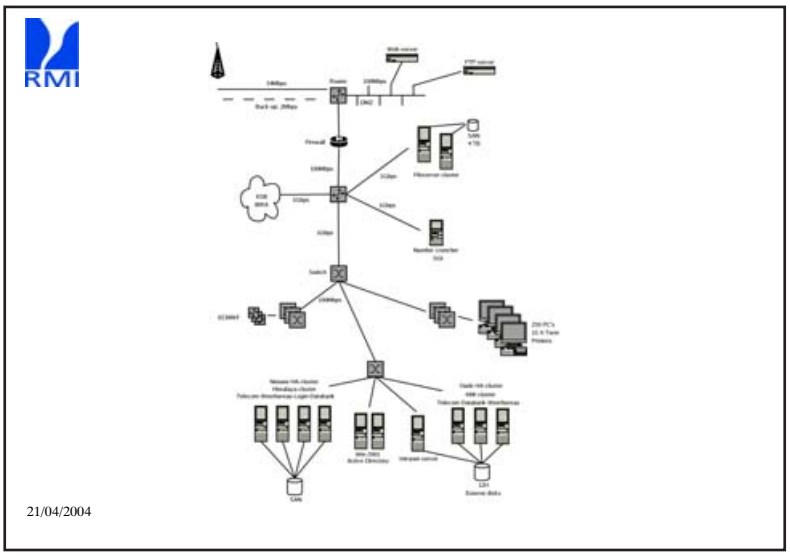
AUSTRIA

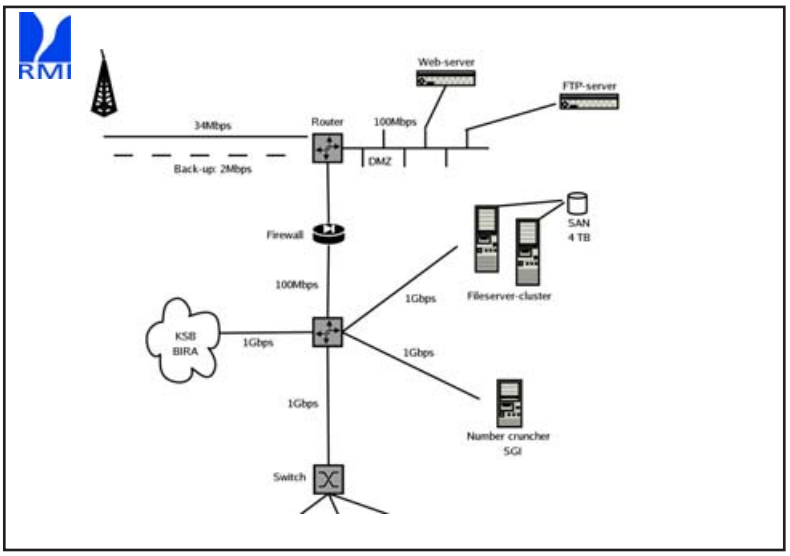


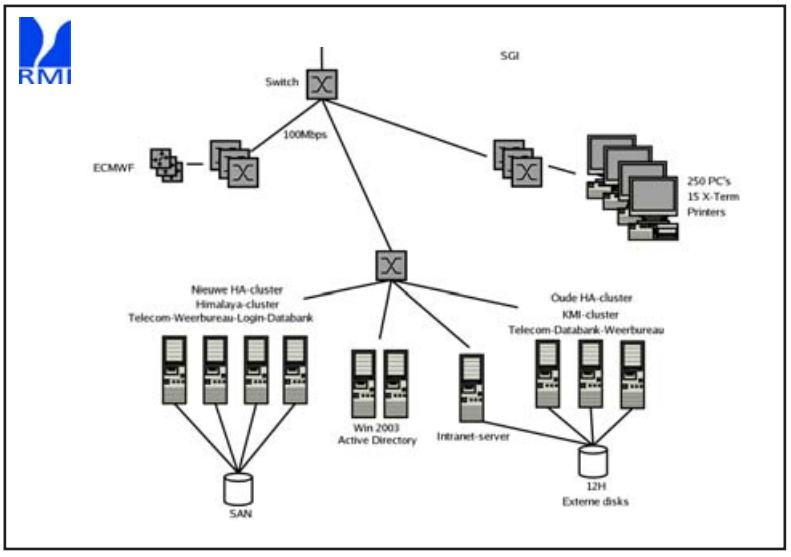
BELGIUM

BELGIUM

Liliane Frappez – Royal Meteorological Institute, Brussels



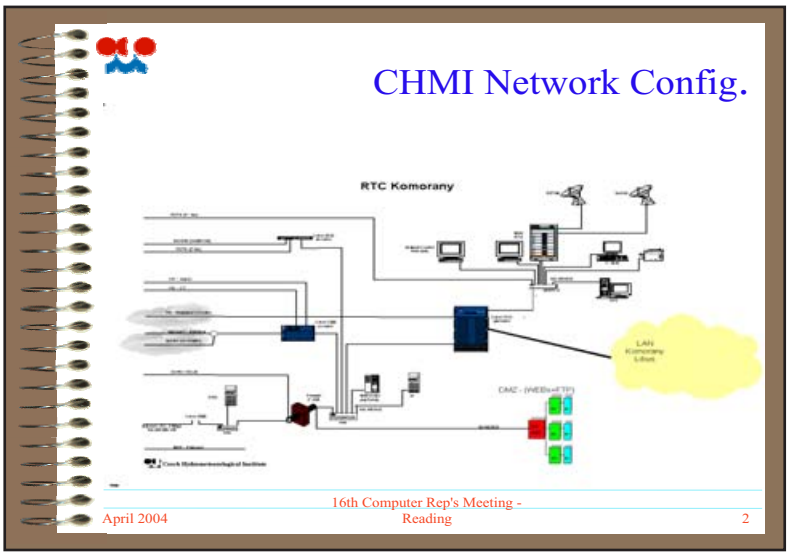


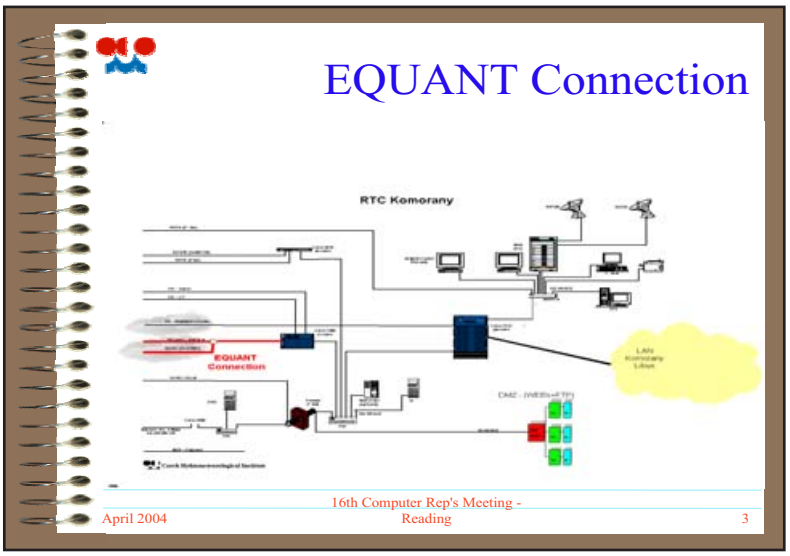


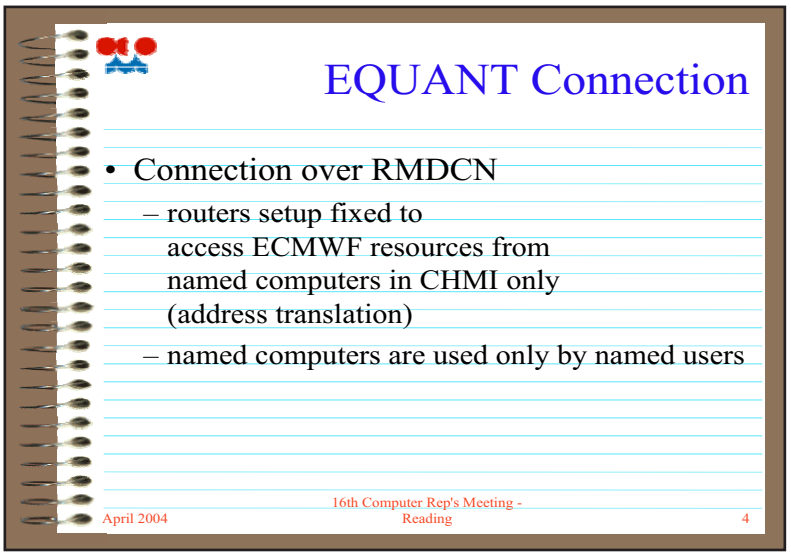
CZECH REPUBLIC

CZECH REPUBLIC

Karel Ostatnicky, Karel Pesata – Czech Hydrometeorological Institute







CZECH REPUBLIC

CZECH REPUBLIC

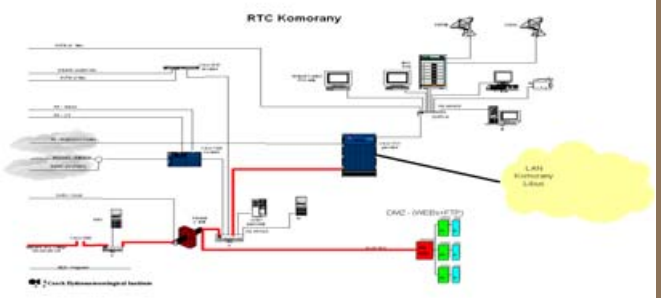
ECMWF Services for CHMI

- *ecgate1* used
- migration to *ecgate* in preparation

ECMWF Services for CHMI

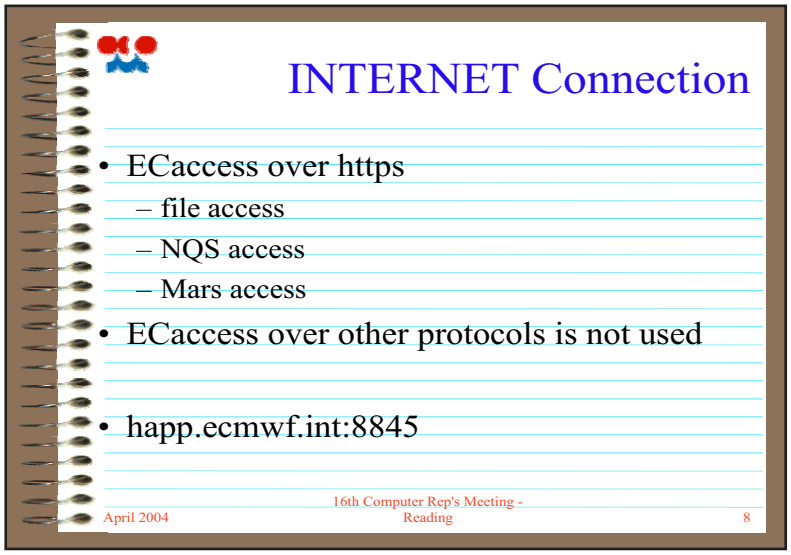
- 5 users
  - 3 CHMI users in meteo-research
    - *Ecgate1* access
    - *https* access
  - 1 CHMI technician
  - 1 user from The Academy of Sciences – Institute of Atmospheric Physics
    - ECMWF data access only

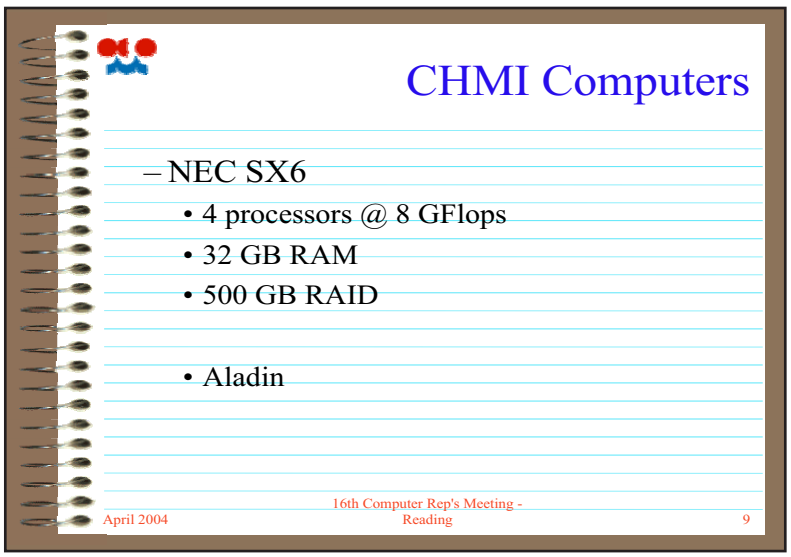
INTERNET Connection

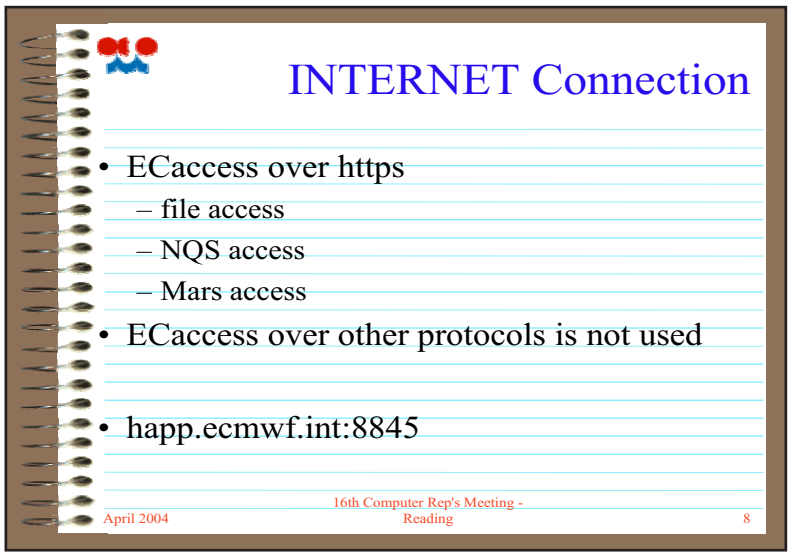


CZECH REPUBLIC

CZECH REPUBLIC



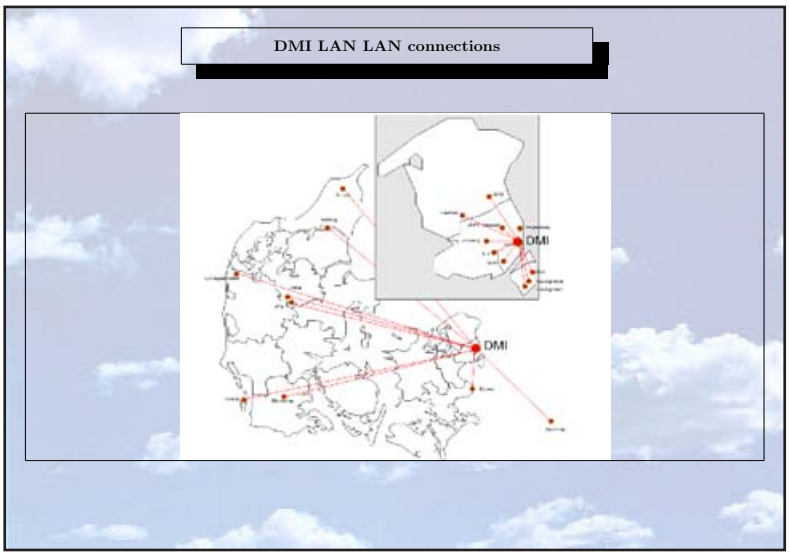


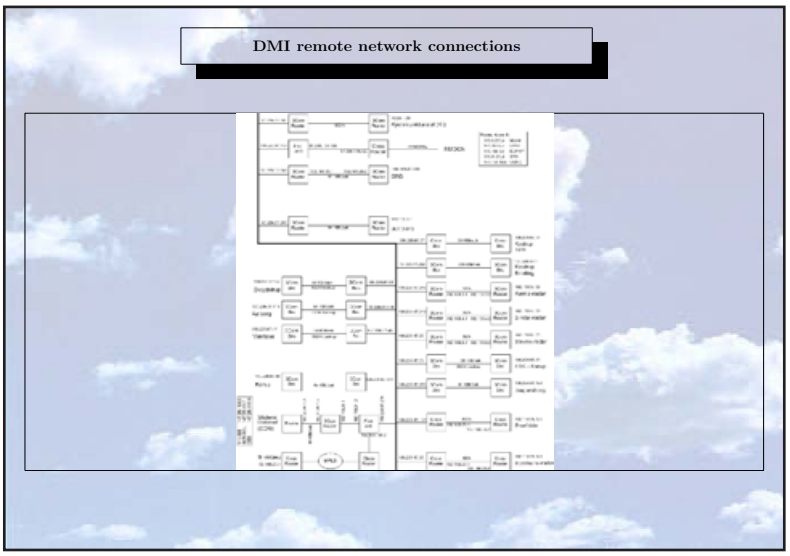


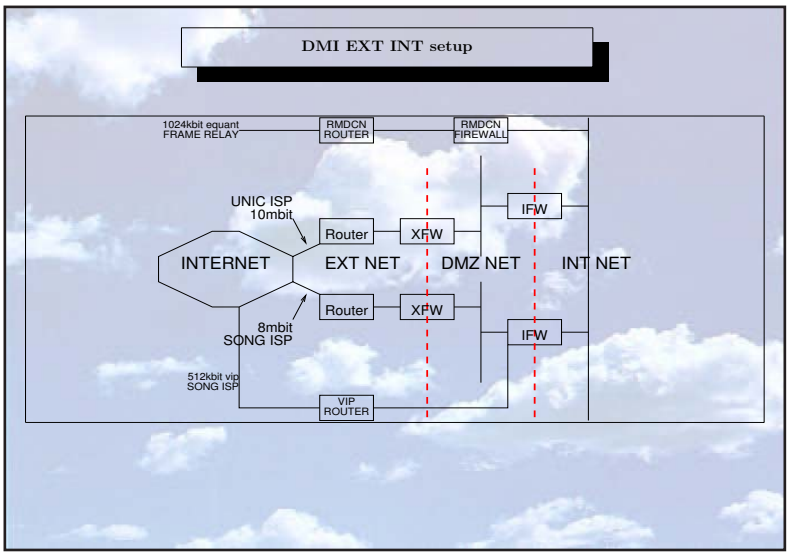
DENMARK

DENMARK

Thomas Lorenzen – Danish Meteorological Institute







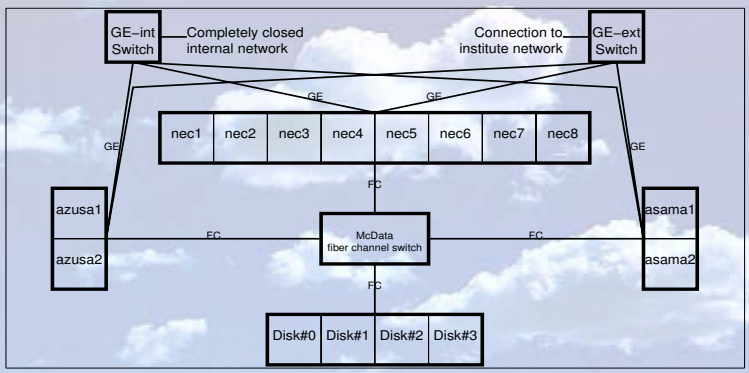
DENMARK

DENMARK

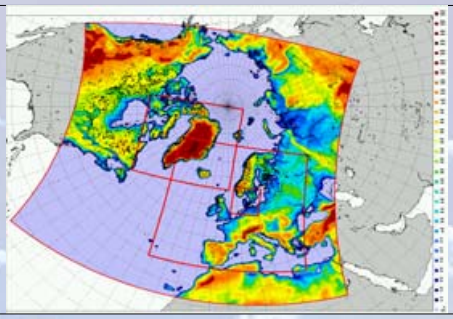
Operational machinepark

- NEC SX6 64M8 (8 vector nodes with 8 CPU each)
  - $64 * 8$ Gflops
  - $32 * 8$ Gbyte RAM
- NEC TX7 16M2 (2 scalar nodes with 8 CPU each)
  - $16 * 1.30$ GHz Intel ItaniumII
  - $16 * 2$ Gbyte RAM
- SX6 is used for running the DMI HIRLAM NWP model four times a day at analysis times 00UTC, 06UTC, 12UTC, 18UTC, whereas TX7 takes care of scalar pre-processing and post-processing.
- SX6 and TX7 share a 4TB disk system via NEC GFS (Global File System), which is a NEC proprietary add on to NFS (Network File System) based on direct access to the 4TB disk system via fiber channel. This eliminates the need for moving data around, which is important, since the amount of data from the DMI HIRLAM NWP model will grow drastically soon.
- To be phased out are 2 4CPU SGI origin200 machines. These 2 4CPU SGI origin200 machines have during the last years served as pre-processing and post-processing machines and have also taken care of the observation decoding.
- Other operational workloads, among which also GTS receiving and archiving from RMDCN, are done on quite a big park of SUN machines, but also LINUX based machines are used.

Supercomputer interconnect



Present forecast areas

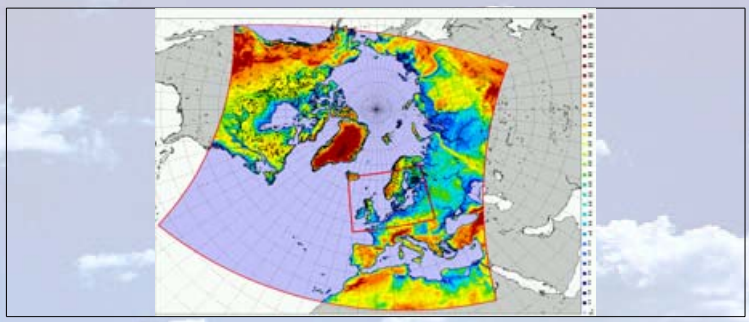


North Atlantic Area : Resolution 450 milli degree on a 202x190x40 grid.
European Area : Resolution 150 milli degree on a 272x282x40 grid.
Greenland Area : Resolution 150 milli degree on a 194x210x40 grid.
Denmark Area : Resolution 50 milli degree on a 182x170x40 grid.

DENMARK

DENMARK

Upcoming forecast areas



North Atlantic Area : Resolution 150 milli degree on a 610x568x40 grid.
European Area : Resolution 50 milli degree on a 496x372x40 grid.

Operational data flow from ECMWF to DMI

- DMI receives from ECMWF via RMDCN data amounting to close to 2GByte per day.
- Of this 2Gbyte amount, more than half of it is frames for the DMI forecast model.
- The remaining data are selected ECWMF forecasts, ensemble forecasts and wave model forecasts, which are plotted for the benefit of the forecasters.
- Around dissemination hours the primary 1Mbit RMDCN line is fully saturated.
- The backup ISDN line only holds one third of the bandwidth of the primary line, so a failure of the primary line will cause operational delays at DMI. The ISDN backup line is not readily upgradeable to 1mbit and is already somewhat expensive.
- An alternative method of getting disseminated data via the Internet has been set up. When the Internet connection is fully functional, transfer rates between ECMWF and DMI are sufficient for not introducing operational delays at DMI.
- DMI currently have 33 registered users using the ECMWF computer systems interactively or via batch jobs via either the ecbatch software or its successor ecaccess, to which users are currently asked to migrate.
- Of the DMI share on the ECMWF supercomputer an amount of 7 percent have currently been used.

FINLAND

FINLAND

Kari Niemelä – Finnish Meteorological Institute

Computing activities at FMI

- No major changes since last meeting
- ECACCESS: comfortable but needs interaction from time to time maybe due to the java version of our gateway server. As a remedy we are planning to transfer the gateway from SGI to Linux
- Anyway the error messages are somewhat misleading and it is difficult to tell what the problem is but…
- Mostly: "eccert error, not authenticated", sometimes "eccert error, not connected" tell the gateway must be restarted
- With the service at ECMWF we are very satisfied. You get an answer very fast when there is something to be asked

ILMATIETEEN LAITOS
METEOROLOGISKA INSTITUTET
FINNISH METEOROLOGICAL INSTITUTE

20/04/2004 1

RCR Hirlam

- The so called reference Hirlam is run on the Finnish IBM testing the new model version.
- 0.2 degree, 438 x 336 points –> lots of data is produced in every run
- ECMWF serves as a distribution centre where the Hirlam community members can pick up the data
- 15 GB of data is transferred daily to ECMWF

ILMATIETEEN LAITOS
METEOROLOGISKA INSTITUTET
FINNISH METEOROLOGICAL INSTITUTE

20/04/2004 2

New organisation

– in effect from 1st March 2004

– strong grouping of similar functions

– 3 large main departments

– vice director general

– commercial services separated

– one new level of superiors

– no removals except the most necessary

– final realisation after removal to new premises

ILMATIETEEN LAITOS
METEOROLOGISKA INSTITUTET
FINNISH METEOROLOGICAL INSTITUTE

20/04/2004 3

FINLAND

FINLAND

New premises

ILMATIETEEN LAITOS
METEOROLOGISKA INSTITUTET
FINNISH METEOROLOGICAL INSTITUTE

20/04/2004 4

New premises

ILMATIETEEN LAITOS
METEOROLOGISKA INSTITUTET
FINNISH METEOROLOGICAL INSTITUTE

20/04/2004 5

New premises

– removal autumn 2005

– FIMR (100) shares the building with FMI (500)

– synergy is expected both from uniting the functions of FMI and from cooperation with FIMR

– a project has been set up to ensure a smooth transfer of functions during the removal

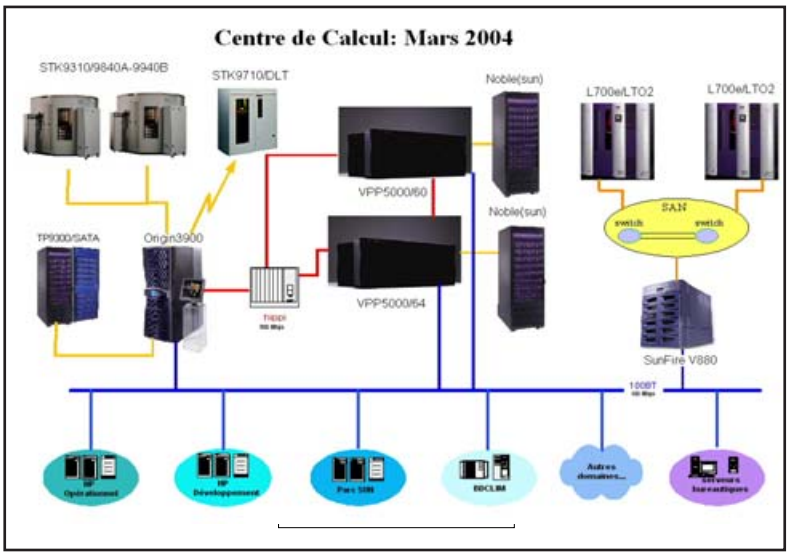
ILMATIETEEN LAITOS
METEOROLOGISKA INSTITUTET
FINNISH METEOROLOGICAL INSTITUTE

20/04/2004 6

FRANCE

FRANCE

Marion Pithon – Météo-France



The compute system

- VPP5000 124 Pes in 2 machines.
- Production : 60 Pes- 280 GB mem- 3 TB disks
- Research,development and backup : 64 Pes- 300 GB mem- 3.9 TB disks
- Office hours maintenance only.
- Production can be switched on the research machine in the event of a failure of the production machine
- Operational files are updated at a regular basis through direct HIPPI link between the 2 VPPs
- Until the end of 2006.

21/04/2004 METEO FRANCE

File server

- New hierarchical storage system installed in March. Software : DMF
- 4 different storage levels :
  - fast Fibre Channel disks (10 TB) for cache
  - Serial ATA disks (20 TB)
  - Fast tapes (9840) in STK 9310 silo
  - Slower tapes (9940) in STK 9310 silo
- Server SGI 03900 (12 procs – 12 GB mem)
- Total capacity 250 TB now, 400 TB at then end of 2004 and 600 TB at the end of 2005.

21/04/2004 METEO FRANCE

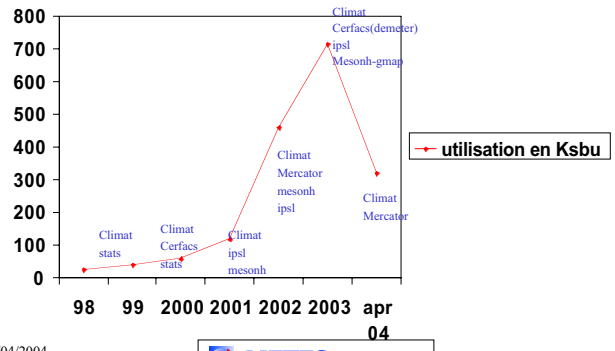
FRANCE FRANCE

Projects at ECMWF

- Operational products : a total of 2 GB received per day
- 45 M.F. projects + 6 Special projects:
  - ✓ 190 users
  - ✓ 136 from Meteo France - 54 from outside
- Main activity is MARS access .
- Feedback from users about use of hpca, migration to ecgate and ECaccess.

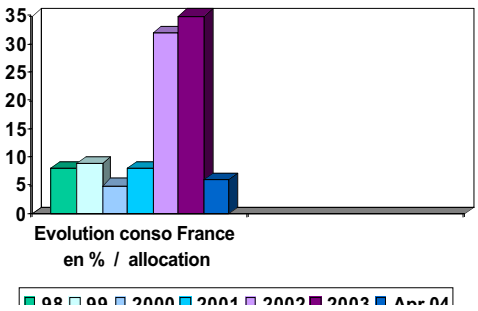
21/04/2004 METEO FRANCE

Billing units use



21/04/2004 METEO FRANCE

French projects HPC resources use (% /alloc)



METEO FRANCE

FRANCE FRANCE

HPCF (1)

- Availability
  - Throughput of jobs is good.
  - Availability is good despite some system or hardware sessions well announced .
- Performance issues
  - Compilations are very fast
  - Optimisation of the codes since last year thanks to ECMWF advices and help :
    - MESONH : efficient help from D. Lucas and P.Tower.
    - MERCATOR : use of MPI_ISEND and choice of data distribution to use a relevant number of nodes and procs/nodes.

21/04/04 METEO FRANCE

HPCF (2)

- Performance issues
  - Important variations of performance (on "MPI only" codes) depending on the number of procs/node and the amount of memory/node :
    - MERCATOR codes run better with 7 procs and 6 GB/node (pagination effect). But some problems when sharing nodes with other codes (code failure).
    - MESONH : best performances when only 4 procs/nodes used
  - In general, CPU costs is higher than on the VPP machine.
  - Users expect significant improvements with the new configuration (new switch and processors).

21/04/04 METEO FRANCE

HPCF (3)

- Specific requirements

  MERCATOR team expresses the need for MPI-2 (spawn of processes and ports mechanism) for their new coupler system PALM-MP
  - IBM's plan about MPI-2 ?
  - Meanwhile, use of LAM (free soft). ECMWF support ?
  - Compatible with ECMWF's environment configuration ?
  - Problem with the ".profile" (trap "0" at the end). A specific ".profile" for MERCATOR users ?

21/04/04 METEO FRANCE

FRANCE FRANCE

Migration to ecgate

- Tests made
  - The majority of users have not started yet.
  - Does not seem to worry them overmuch.
  - Better performances with compilations.
- Specific requirements :
  - Need support from user support for specific software
  - Help of Dominique Lucas for NCAR installation (MESONH)
  - What are ECMWF plans about CVS on ecgate ? Installation of new versions ?
  - When will SMS jobs run on ecgate ? Automatic migration ? Users have to modify their jobs (#QSUB parameters) ?
  - What about cron jobs ? Migration made by users when they are ready ?

21/04/04 METEO FRANCE

ECaccess (1)

- Administration
  - Has made the use of Internet easier between ECMWF and Meteo France.
  - Version of gateway : 2.1 (upgrade easy with the help of L.Gougeon)
  - 16 destinations for ectrans. 52 MS users registered.
  - 2 other gateways in institutes outside M.F
- Connections
  - Easier X11 access then before
  - Experiences of slow response time for telnet (slower then through RMDCN). Packet shaper configuration at Météo France ?

21/04/2004 METEO FRANCE

ECaccess (2)

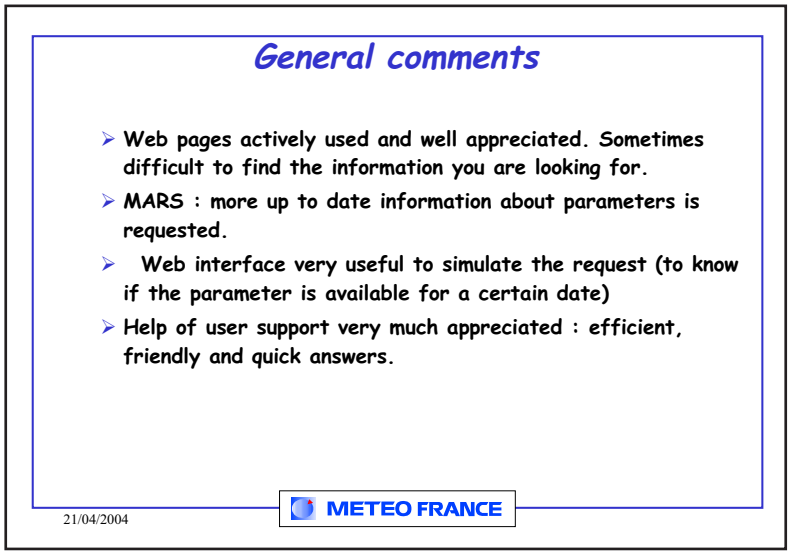
- File transfers

Users are enthusiastic about ectrans command

- Easy to use
- Convenient for batch jobs
- Reliable
- Very flexible and convenient (choice of target directory, possibility to create the target directory if does not exist, …)
- Warning by email on failure very useful
- Faster rate for transfers (internet bandwidth better than RMDCN)

21/04/2004 METEO FRANCE

FRANCE FRANCE



GERMANY GERMANY

Elisabeth Krenzien – Deutscher Wetterdienst

Deutscher Wetterdienst

Contents

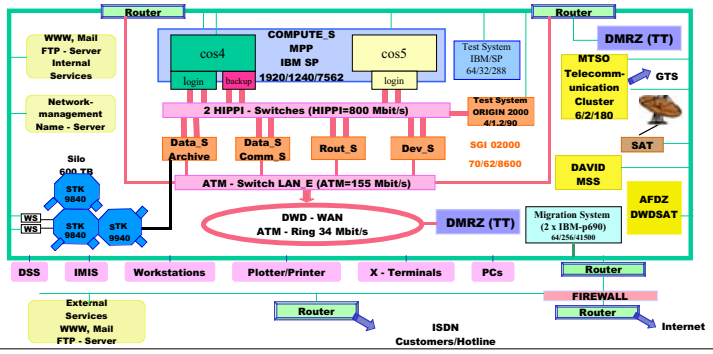
- Computing infrastructure
  - IBM SP Compute server upgrade
  - Replacement of central SGI O2000 systems
  - Connections to ECMWF
  - Plans
- Projects at ECMWF: Status and Plans
  - DMRZ Operations
  - Research modelling
  - Special Projects
- Experiences of users

16th CR Meeting, ECMWF, 2004 - 2 - DWD - April 2004

Deutscher Wetterdienst

Configuration 2004 - DMRZ (OF)

Status: March 2004 Keys: number processors/memory in GB/disc space in GB

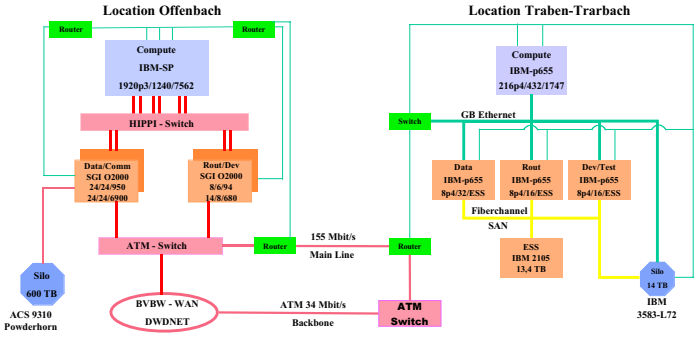


16th CR Meeting, ECMWF, 2004 - 3 - DWD - April 2004

Deutscher Wetterdienst

German Meteorological Computing Center (DMRZ)

Status: March 2004 Keys: number processors/memory in GB/disc space in GB



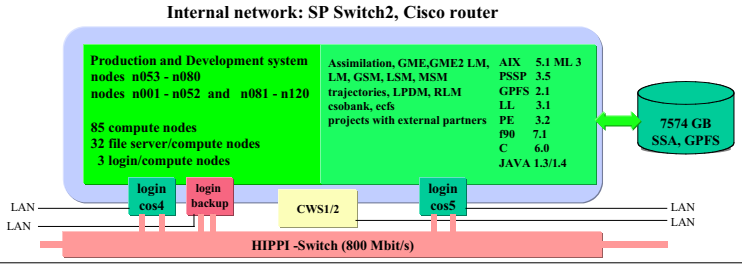
16th CR Meeting, ECMWF, 2004 - 4 - DWD - April 2004

GERMANY GERMANY

Deutscher Wetterdienst

Compute server: upgrade (1)

- Addition of 40 16-way NH II nodes, 3.8 TB SSA disk space
- Model of operation:
  - single system with one instance of GPFS and LoadLeveler;
  - development and production separated by definitions of LL job classes



16th CR Meeting, ECMWF, 2004 - 5 - DWD - April 2004

Deutscher Wetterdienst

Compute server: Migration from AIX 4.3.3 to AIX 5.1 (1)

- **Reason for migration:**
  - End of support for AIX 4.3.3 on RS/6000 SP in 2004
  - Operational Suite will expand to use almost all 120 nodes
  - Suspending (user) jobs requires LL 3.1 features (preemption in gang sched.)
- **Conditions and requirements**
  - Hard restriction to a time slot of 4 hours per day to garantee production
  - Test system with new AIX 5.1 image
  - Backup CWS with Aix 5.1 for the installation on the nodes
  - unmirroring of rootvg
  - 2 installation runs for 8 nodes in a maintenance window

16th CR Meeting, ECMWF, 2004 - 6 - DWD - April 2004

Deutscher Wetterdienst

Compute server: Migration from AIX 4.3.3 to AIX 5.1 (2)

- **Summary**

  - *November, 4th*: Boot of AIX 5.1 on whole system and VSD/GPFS mirroring AIX 5.1 rootvg on all nodes and upgrade of primary CWS

  - Migration to AIX 5.1 including planning, testing and installation took nearly six months
  - Test system is an important assumption for the migration
  - Restriction to 4 hour maintenance time slot on the production system could be fulfilled
  - Excellent preparation and support during the migration by IBM Software Support

16th CR Meeting, ECMWF, 2004 - 7 - DWD - April 2004

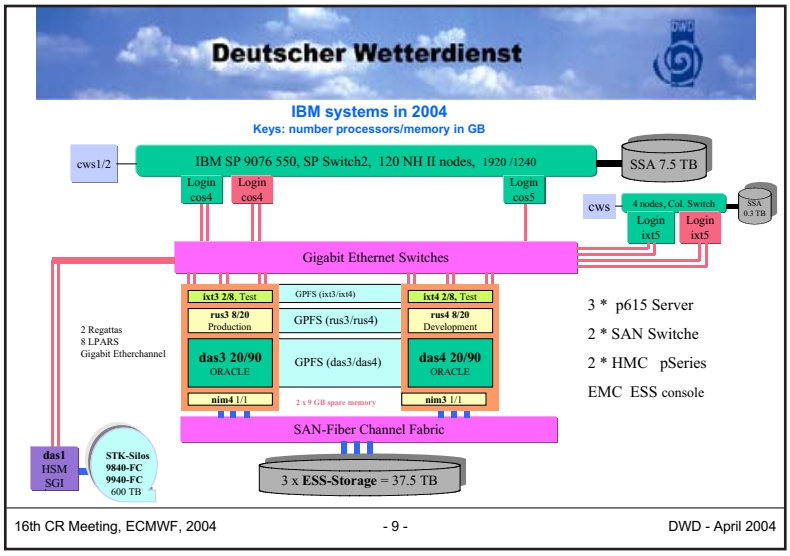
Deutscher Wetterdienst

Replacement of SGI O2000 systems

- November 2003: delivery of two p690 systems
  - 32 CPUs, 1.7 GHz Power4, 128 GB each
  - 3 * ESS with 12.5 TB disk capacity each
  - 3 * p615 systems as quorum servers for GPFS on ESS and p690 LPARs
  - 2 *HMCs
- December 2003: Acceptance test passed
  - two p690 systems are divided symmetrically into 2*4 LPARs
  - data handling server LPARs (20 CPUs, 90 GB, HACMP cluster)
  - operational pre/post processing, user login server LPAR (8 CPUs, 20 GB)
  - test system LPARs (2 CPUs, 8GB, HACMP cluster)
  - NIM server LPARs (1 CPU, 1 GB)

16th CR Meeting, ECMWF, 2004 - 8 - DWD - April 2004

Deutscher Wetterdienst



16th CR Meeting, ECMWF, 2004 - 9 - DWD - April 2004

Deutscher Wetterdienst

Plans and Challenges

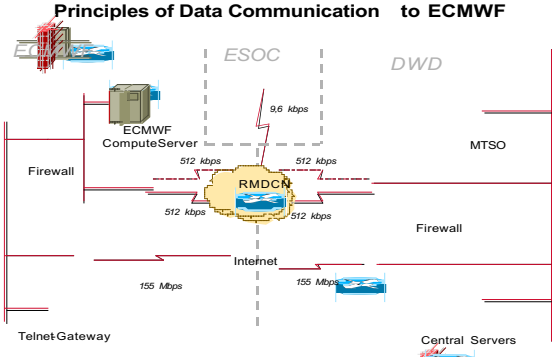
- **Replacement and Migration**
  - SGI O2000 servers for Database Management to IBM p690 systems
  - SGI O2000 servers for meteorological products and for development to IBM p690 systems
  - User environment (~ 700 users)
  - **power off : 1st July 2004**
- **Replacement of Archiving system (server and HSM)**
  - Proof of Concept currently in progress
  - ITT in late summer
- **Kick off planning of next compute server in 2007**
  - Definition of next generation forecast suite's computing resources
  - Negotiate funding

16th CR Meeting, ECMWF, 2004 - 10 - DWD - April 2004

GERMANY GERMANY

Deutscher Wetterdienst

Principles of Data Communication to ECMWF



16th CR Meeting, ECMWF, 2004 - 11 - DWD - April 2004

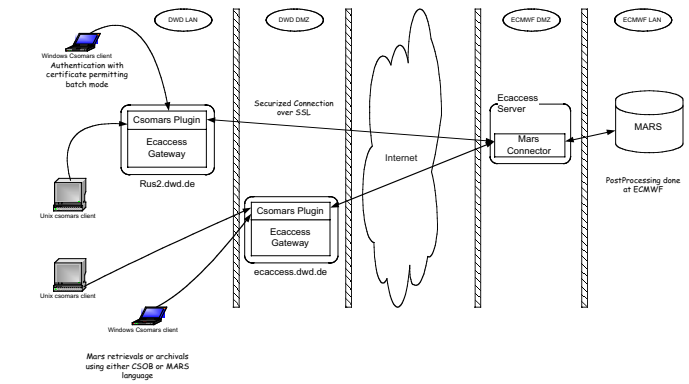
Deutscher Wetterdienst

Projects: DMRZ Operations (1)

ECFS enhancements:	**AIX 5.1 clients**
	introduction of ecquotas
	development of MS_NT client (ECcmd based)
ECaccess gateways:	**Linux (DMZ), Irix (LAN), planned (RMDCN)**
Csomars:	**version 1.2 in full operation**
	remote access to MARS, interactive client
SMS:	**single instance solely for operational production**
UNICORE 4.1:	**beta testing**

16th CR Meeting, ECMWF, 2004 - 12 - DWD - April 2004

Deutscher Wetterdienst



16th CR Meeting, ECMWF, 2004 - 13 - DWD - April 2004

GERMANY

GERMANY

Deutscher Wetterdienst

Projects: Research Modelling (2)

GME reference	**successful for operation at DWD** **improvements for assimilation schemes**
Muse Project	**case study of storm surges in the North Sea (5 episodes)** **identify members in EPS/IFS experiments, data transfer** **LM runs at DWD; *end in summer 2004***
LME testbed	**in place since March, one run per day**
ODB SW package	**installed at DWD, filling in test data in progress** **test applications for 1D and 3 D Var assimilation**

16th CR Meeting, ECMWF, 2004 - 14 - DWD - April 2004

Deutscher Wetterdienst

Research jobs on HPC systems (1)

Global Model GME

Icosahedral hexagonal gridpoint model (Majewski et al., MWR, 2002)

***Goals*: Test of new horizontal and vertical resolutions; perform daily runs from 12 UTC with the IFS (T_L511/L60) initial state.**

***Jobs*: *ifs2gme* – Interpolation of the IFS-analysis to the GME grid**
***gmtri* – GME model**

***Runs*: GME 60 km, L40, 7 x 8 Proc., 174 h, 373 BU**
GME 40 km, L40, 7 x 15 Proc., 174 h, 1225 BU
Forecast data stored in MARS (*dwd2mars*): 20.7 GByte/day

***status*: continue**

16th CR Meeting, ECMWF, 2004 - 15 - DWD - April 2004

Deutscher Wetterdienst

Research jobs on HPC systems (2)

IFS2GME

Interpolation of IFS data to the GME grid (60 km and 40 km)

***Goals*: Provide ozone analysis (00 UTC) for the operational GME data assimilation and forecast system as a basis for UV-B prediction. Provide stratospheric temperature, water vapour, ozone profiles for experimental 1D-Var assimilation of ATOVS (NOAA) data for 00, 03, 06, …, 18, 21 UTC.**

***Jobs*: *ifs2gme* – Interpolation of IFS-analysis to GME grid**

***Status*: continue**

16th CR Meeting, ECMWF, 2004 - 16 - DWD - April 2004

GERMANY

GERMANY

Deutscher Wetterdienst

Research jobs on HPC systems (3)

"Pseudo"-Temps

Interpolation of IFS fields to the reduced lat/lon grid (190km mesh)

***Goals*: Derive "Pseudo"-Temps from the IFS-analysis 00 UTC over the oceans and Antarctica for the operational GME assimilation analysis at 00 UTC (performed at 12.30 UTC).**

***Jobs*: *csomars;* BUFRs created at the DWD**

***Impact:* Substantial improvement of the GME analysis and forecast**
SH: gain of up to 24 h at day 5; NH: a gain of up to 6 h at day 5

***Status:* continue**

16th CR Meeting, ECMWF, 2004 - 17 - DWD - April 2004

Deutscher Wetterdienst

Local Model LME

Nonhydrostatic regional model for whole of Europe (Doms et al., DWD, 2003).

Mesh:	**7 km, L40; 665 x 657 gridpoints; time step: 40 s; 12 UTC + 72 h**
***Initial state*:**	**IFS analysis**
***Lat. bound. data*:**	**GME 40 km, L40, 3-hourly, later on hourly**
***Goals*:**	**Sensitivity tests of LME (numerical schemes, parameterizations)**
***Jobs*:**	***gme2lm* – Interpolation of the GME data to the LME grid** ***lm* – LME model**
***Runs*:**	**gme2lm, 4 x 8 Proc., 72 h, 23 BU** **lm, 10 x 14 Proc., 72 h, 1745 BU** **Forecast data stored in MARS (*dwd2mars*): 12 GByte/day** ***Initial data*: 300 MB, *lateral boundary* data: 6 GB (in ectmp:)**
***Status*:**	**one run per day since 30 March 2004 ; up to 3 runs in the future**

16th CR Meeting, ECMWF, 2004 - 18 - DWD - April 2004

Deutscher Wetterdienst

Projects: Research Modelling (3)

ELDAS studies	**European Land Data Assimitation System** **DWD: soil moisture assimilation** **will be implimented at DWD**
New Assimilation Project	**national funded, centered on Data assimilation** ***global scale*: Ensemble transpose Kalman filtering** ***convective scale:* LME based**
CM_SAF	**Satellite Application Facility for Climate Monitoring** **evaluation phase for 'pre-operation' or 'repository'** **kick-off (?) visit in May**

16th CR Meeting, ECMWF, 2004 - 19 - DWD - April 2004

GERMANY

GERMANY

Deutscher Wetterdienst

Special Projects

MPI, Hamburg: early tests on new IBM System (ECHAM5 validation)
tests for MOZART2.1, beta-testing ecgate
installation of PRISOM environment

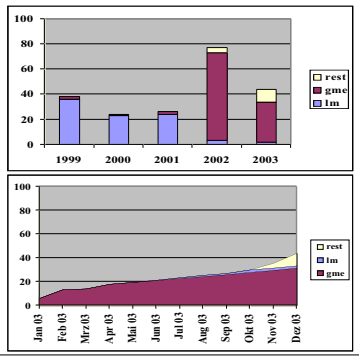
DLR, Oberpfaffenhofen:
influence of non-hydrostatic gravity waves on stratospheric flow over mountains
current forecast data for mesoscale field campaigns

1 Special Projects: Mars Data Retrieval

16th CR Meeting, ECMWF, 2004 - 20 - DWD - April 2004

Deutscher Wetterdienst

HPCF Allocations (in %)



Statistics

Total number of users:	2004		2003
DWD:	64 (datex: 20)	25 (April) 44 (2004)	63
SP:	66	19 (April) 42 (2004)	54

Usage of storage

2003: 19.8 TB available
6.7 TB in ECFS 10.5 TB in MARS

2004: 41.8 TB available
6.9 TB in ECFS 12.1 TB in MARS

16th CR Meeting, ECMWF, 2004 - 21 - DWD - April 2004

Deutscher Wetterdienst

Experiences

Users appreciate the professional support from ECMWF Staff, especially from Norbert (User Support) and Petra (SecureID cards) and from Research

Concerns raised are	**IBM Compiler versioning, shortage of disk space, 2 GB file size in ECFS, MPI2 support (mpi_comm_spwan) (MPI, Hamburg) reliability of MS jobs (disappearing nqe jobs, delays)**
questions refer to	**accessing cvs servers in MS functionality on ecgate/ecgate1**

16th CR Meeting, ECMWF, 2004 - 22 - DWD - April 2004

Deutscher Wetterdienst

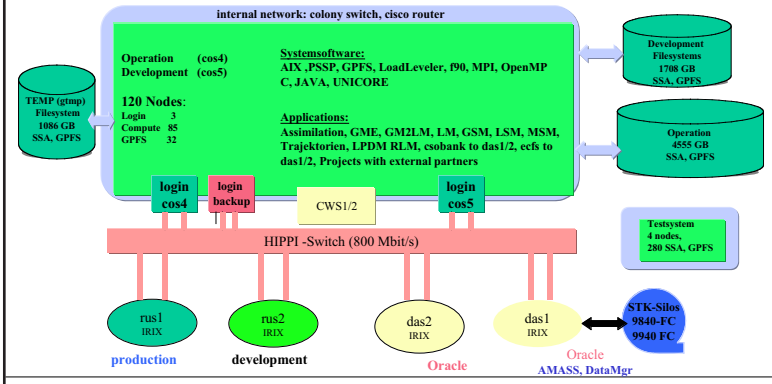
Projects: DMRZ Operations (2)

ODB:	AIX 5.1 clients
	introduction of ecquotas
	development of MS_NT client
CM_SAF:	Linux (DMZ), Irix (LAN), planned (RMDCN)
Csomars:	version 1.2 in full operation
	remote access to MARS, interactive client

16th CR Meeting, ECMWF, 2004 - 24 - DWD - April 2004

Deutscher Wetterdienst

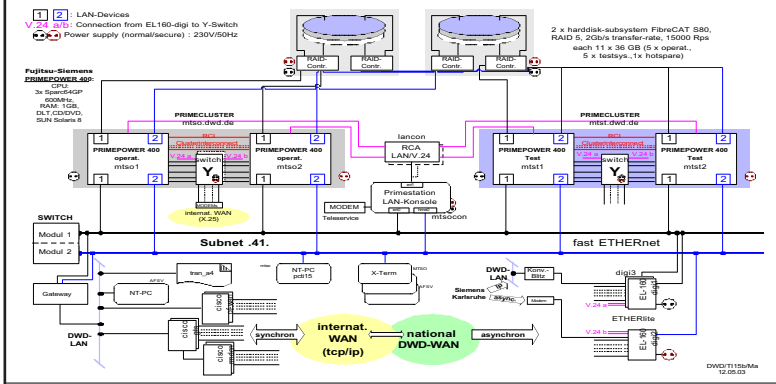
IBM SP 9076 550 (NH II, 375 MHZ)



16th CR Meeting, ECMWF, 2004 - 25 - DWD - April 2004

Deutscher Wetterdienst

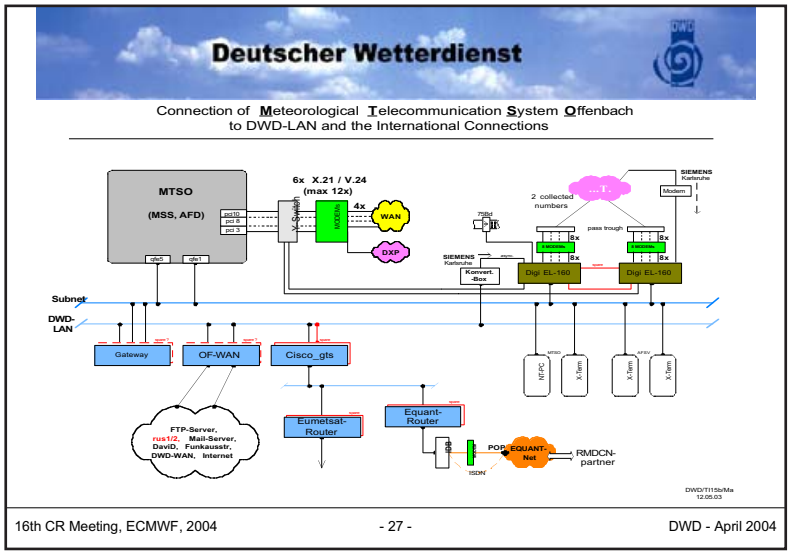
MTSO - Meteorological Telecommunications System Offenbach



16th CR Meeting, ECMWF, 2004 - 26 - DWD - April 2004

GERMANY

GERMANY



GREECE

GREECE

Ioannis Mallas – Hellenic National Met. Service

Hellenic National Meteorological Service

Modernization plan of HNMS

Goals

- Meteorological support of the Athens 2004 Olympic Games.
- To upgrade the observation network and processing systems in order to provide more accurate weather products orientated to customer and internal end-user needs.
- Improve HNMS support to Hellenic early warning system for hazardous weather phenomena.

Hellenic National Meteorological Service

Modernization plan of HNMS

Status of Computer and System Infrastructure

Main Computer for the run of HNMS NWP Models	OK
C- Band Doppler Radar (Aigina Island)	OK
MSG Satellite Station.	Installation Phase
Now-Casting System.	Testing Phase
Meteorological Visualization W/S System.	Testing Phase
Meteorological Archival and Retrieval System (MARS)	Installation Phase
Preprocessing System (ECMWF)	Testing Phase

Cont. …

Hellenic National Meteorological Service

Modernization plan of HNMS

Status of Computer and System Infrastructure

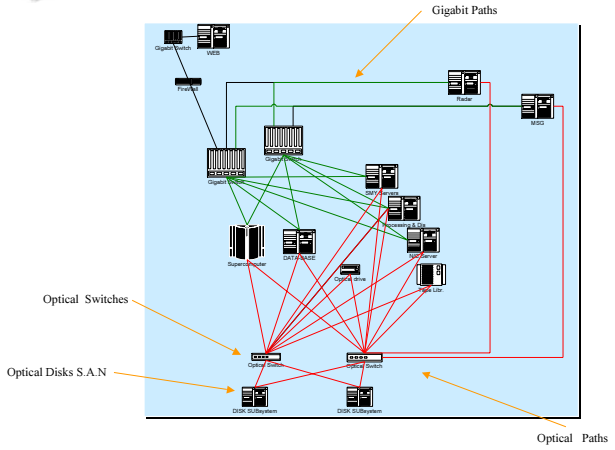
MSS	Testing Phase
SMS	Testing Phase
WEB Server	Installation Phase
Main Super Computer	Benchmarking Phase
Lightning Detection Network and Wind Profiler Radar.	ITT
Automatic – Manned Weather Station Network Upgrade	ITT

GREECE

GREECE

Hellenic National Meteorological Service

Planned Main Computer Facilities



Hellenic National Meteorological Service

NWP Super Computer (Interim)

HP RX2600 - 28 Nodes System
- 64 bit OS. HPUX 11.22
- Mem 80GB
- Link Interconnection >= 1GB/sec
- DISK SubSystem
  - SAN Connectivity
  - RAID Support
  - 100MB/sec
  - 2 Tbytes / 600 GB

- MPI
- RUNs LM

Hellenic National Meteorological Service

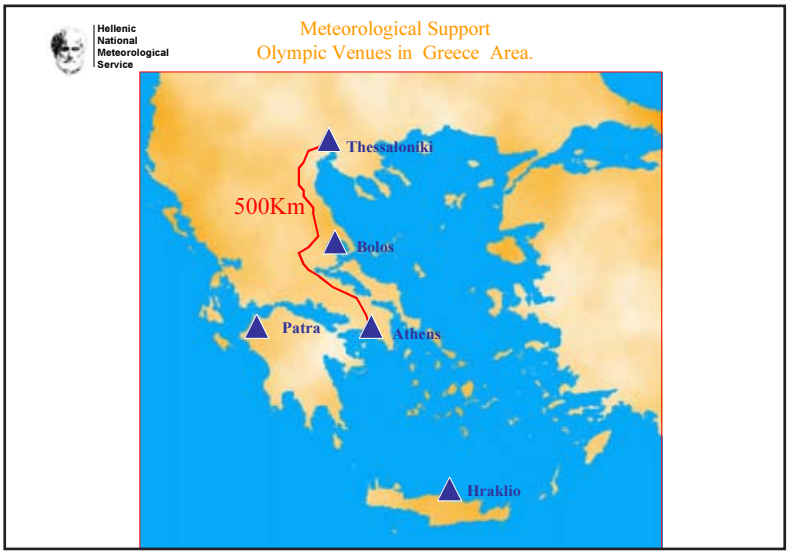
MARS - WEB MARS

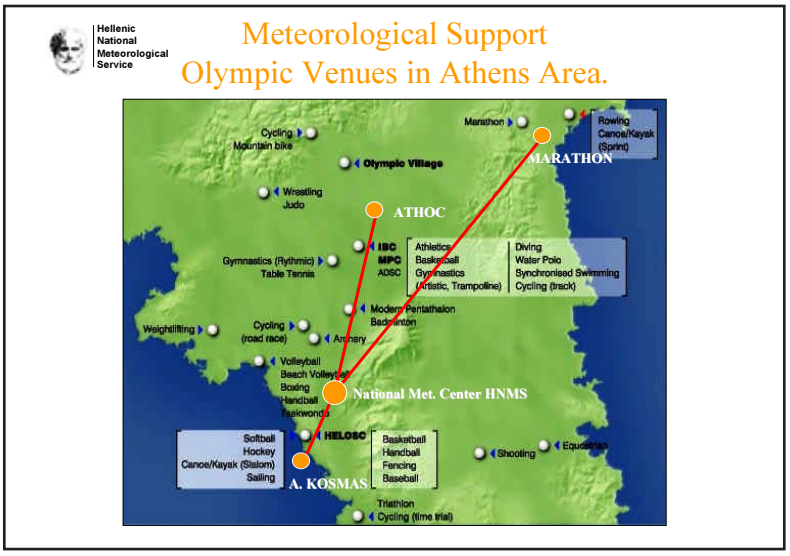
- HP RX5670 High Availability Server (2)
- OPTICAL DISK SubSystem
  - SAN
  - RAID Support
  - 100MB/sec
  - 11 Tbytes (ext.)
- Robotic Library
  - 13TBytes (ext.)

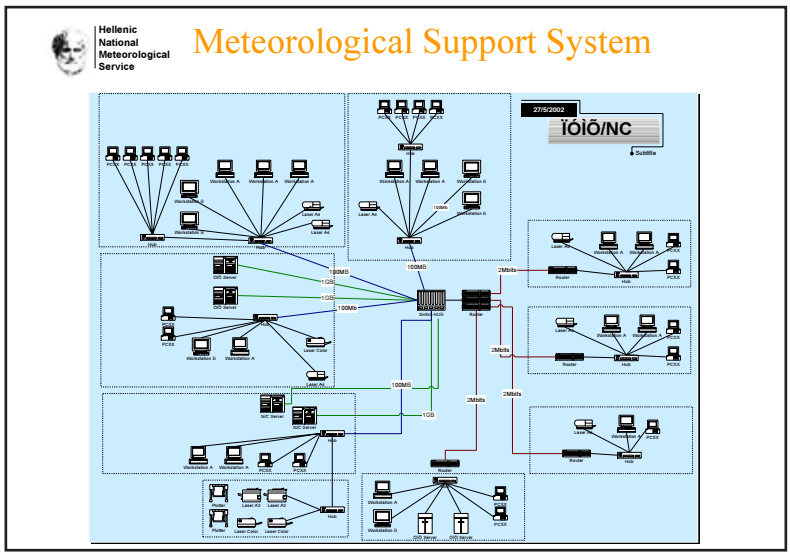
- MARS Server
- MARS Client
- WEB MARS

GREECE

GREECE

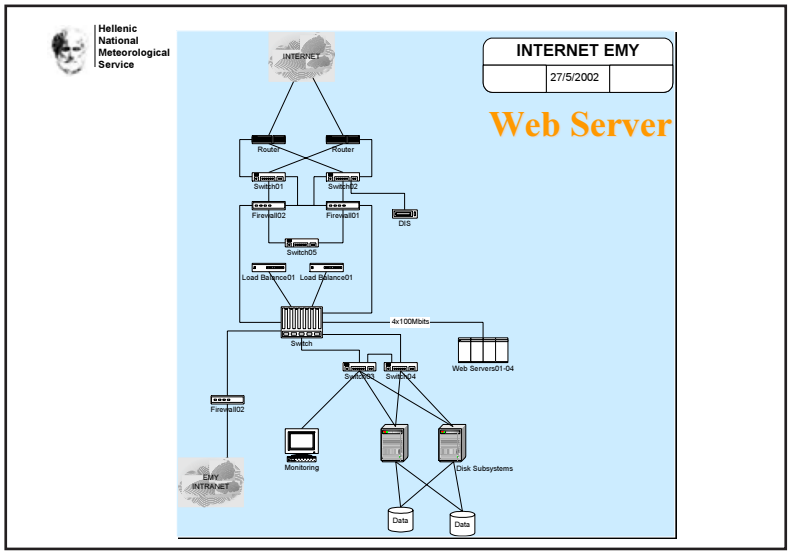


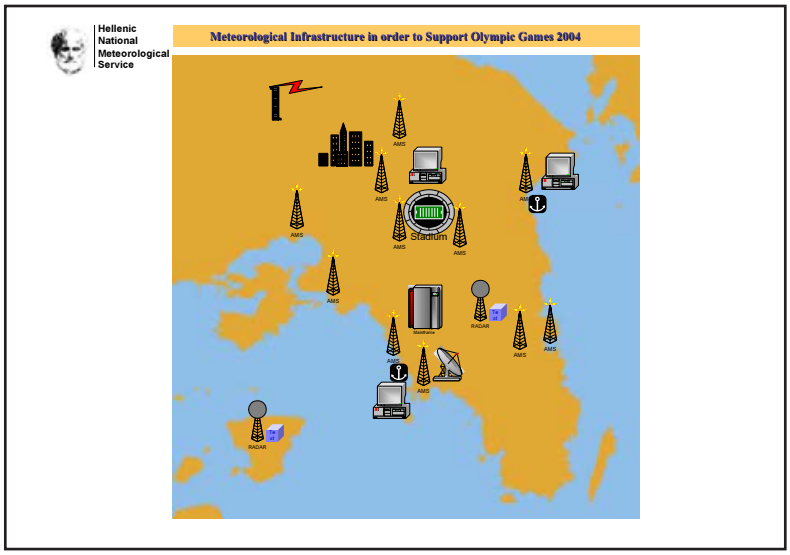




GREECE

GREECE





Hellenic National Meteorological Service

Use of ECMWF Facilities

- Deterministic Model
- EPS
- Wave Model
- Boundary Condition
- Daily run of LM
- Trajectory Model
- MARS Retrieval

HUNGARY

HUNGARY

László Tölgyesi – Hungarian Meteorological Service

Network

- LAN: 100/1000 Mb/s, structured UTP cabling system, CISCO switches
- SAN: 2Gb/s, redundant FC-switches, CLARiiON connected: IBM, HPcluster, SGI, Linux
- WAN: 64 kb/s and 128 kb/s leased lines, 2 Mb/s and 4*2 Mb/s microwave channels
- Internet: 512 kb/s and 128 kb/s
- RMDCN (256 kb/s with ISDN backup, upgraded from 192 kb/s):
  - ECMWF: 128 in / 16 out kb/s;
  - Austria (AC RTH): 32 in / 8 out kb/s
  - Slovakia: 16 in / 16 out kb/s
- Satellite broadcast: SADIS, RETIM 2000
- Firewall (old:Gauntlet; new:ZORP): Internet segment and DMZ

2 Report on the 16th meeting of Member State Computing Representatives, 21-22 April 2004, Hungary

Computer resources

- Process & Application Server I. (IBM pSeries 690; 32 CPU, 64 GB RAM, 850 GB disk, SAN): LAM Aladin-Hu, other research studies
- Process & Application Server II. (SGI ORIGIN 2000; 16 CPU, 8 GB RAM, 88 GB disk, SAN): nowcasting modelling (MM5), GIS, WEB
- Central Processing & Controlling System (HP L3000 cluster-PKG2; 4 CPU, 3 GB RAM, SAN + HP D280, K250): scripts, programs; CDS-CASS
- Database server (HP L3000 cluster-PKG1; 4 CPU, 3 GB RAM, SAN): ORACLE (8.1.6), CLDB, CADB
- Message Switching System (2 PC-s; Linux): life-standby WeatherMan
- Application Server III. (PC Linux cluster with 4 nodes; Pentium IV. 2.4 GHz CPU, 0.5 GB RAM, 40 GB HD; SAN): WEB (*test phase*)
- Central Storage System: (CLARiiON FC4700; ~2.9 TB, with HP SureStore Ultrium 2/20 for system backup and HP DLT 1/8 for data backup)
- Other (firewall, mail, printer, WAP, WEB…) servers: Linux, Unix, Netware
- DEC, SUN, HP and Linux WS's for visualisation and development
- about 300 PC (Windows, Linux)

3 Report on the 16th meeting of Member State Computing Representatives, 21-22 April 2004, Hungary

Cluster Project

Operative configuration (*test phase*)

PC Linux cluster with 4 nodes (1 master, 3 computing)
Pentium IV(Xeon), 2.4 GHz CPU, 0.5 GB RAM, 40 GB HD per node
GigaBit CISCO switch amoung nodes (1 Gb/s; UTP)
SAN (2 Gb/s), LAN (1 Gb/s) connected to master node
OS: Linux 2.4.20
Cluster SW: OSCAR (*Open Source Cluster Application Resources*)
Loadbalancing, scheduling: Maui (*tested on WEB*)
Job controlling: OpenPBS (*Portable Batch System*)

Note: *SGI and HP machines with two Itanium CPU are tested*

Planted configuration (*December 2004*)

PC Linux cluster with 17 nodes (1 master, 16 computing)
Two Pentium IV(Xeon), 2.4 GHz CPUs, 2 GB RAM per node

…and later

Linux cluster with 32 or 48 nodes
Itanium 2, 1.5 GHz CPUs, 2 GB RAM per node

4 Report on the 16th meeting of Member State Computing Representatives, 21-22 April 2004, Hungary

HUNGARY

HUNGARY

Changes related to ECMWF

- Sixteen registered users since April 2004 (ten users were in 2003)
- Operational use of 0.5 x 0.5 degrees deterministic forecast and 1.0 x 1.0 degrees ensemble forecast twice a day for European area (since January 2004)
- Operational use of EPS clusters for Central European area and ectrans for file transfer from ecgate to HMS server via Internet (since July 2003)
- Installation of the newest version of ecaccess software was done last November.
- WEB based application management: NWP monitor
- Local questionnaire on use of ECMWF resources (July 2003 and March 2004)
- Local training course on ecaccess and migration to ecgate (9-11 March 2004)
- Welcome early delivery system (16 March 2004)

- No projects run at ECMWF

5 Report on the 16th meeting of Member State Computing Representatives, 21-22 April 2004, Hungary

Summary of questionnaire on use of ECMWF resources

Q.1: computer usage
50 % work on both ecgate1/ecgate and local computer
40 % work on only ecgate1/ecgate
10 % work on only local computer

Q.2: file transfer mode
70 % use ectrans or ecput/ecget
30 % use ftp

Q.3: type of work on ecgate1/ecgate
50 % operational and research & development (R&D)
30 % only R&D
10 % only operational
10 % don't work on it

Q.4: migration to ecgate
40 % have been started and experiences are promising
60 % are aware of deadline (31 July 2004) but hasn't started

6 Report on the 16th meeting of Member State Computing Representatives, 21-22 April 2004, Hungary

Summary of questionnaire on use of ECMWF resources (cont)

Q.5: MARS and user's guide usage
100 % use MARS
50 % only on ecgate1/ecgate
70 % use web information
50 % use printed documents

Q.6: MAGICS and user's guide usage
40 % use MAGICS
30 % only on ecgate1/ecgate
100 % of them uses web information and printed documents too

Q.7: METVIEW and user's guide usage
40 % use METVIEW
30 % only on ecgate1/ecgate
100 % of them use web information
75 % of them use printed documents
75 % of them ask Comp. Representative or User Support if it's needed

7 Report on the 16th meeting of Member State Computing Representatives, 21-22 April 2004, Hungary

HUNGARY

HUNGARY

Summary of questionnaire on use of ECMWF resources (cont)

Q.8: Trouble shouting
80 % use ECMWF web
80 % ask Computing Representative
70 % read printed documents, 60 % ask colleagues,
40 % occasionally ask User Support

Q.9: Quality of printed documents
40 % said: good, clear and well organised
60 % said: suitable

Q.10: Quality of ECMWF web information
80 % said: good, clear and well organised
20 % said: suitable

Q.11: Assistance of Computing Representative
All of them is satisfied

Q.12: Assistance of User Support
All of them is satisfied

8 Report on the 16th meeting of Member State Computing Representatives, 21-22 April 2004, Hungary

Summary of questionnaire on use of ECMWF resources (cont)

Q.13: Need of additional information and/or training *(more answers)*
60 % local training courses
20 % ECMWF training courses
20 % more information on ECMWF web
10 % printed documentation
30 % don't know the future needs,
20 % have no additional needs

Q.14: Subject of local training course on ecgate/ecaccess and advanced MARS (9-11 March 2004)
70 % were fully satisfied,
20 % said: training was good and it was just enough,
10 % said: it was timely and necessary but it was a bit too much

Q.15: Experiences of local training course
90 % said: it was easy to follow
10 % said: more practical information should have been provided

9 Report on the 16th meeting of Member State Computing Representatives, 21-22 April 2004, Hungary

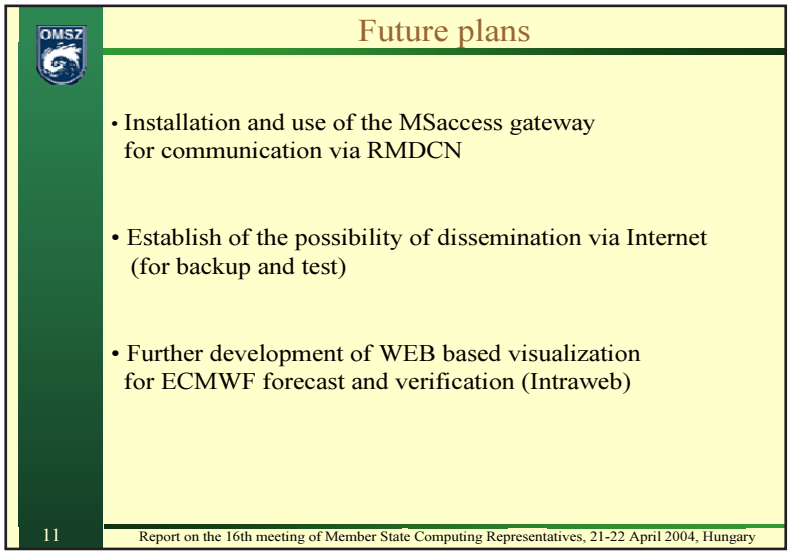
ECMWF data coming to HMS

Data type	number of files	MB/day	arriving time [UTC]
DATA COMING VIA DISSEMINATION			
European area (70N,15W, 34N,40E; DET: 0.5x0.5, EPS: 1x1 degrees)			
H2D - GRIB DET 00&12 UTC	53	2*223	10.00 pm/am - 1.00 am/pm
H2E - GRIB EPS 00&12 UTC	49	2*246	1.00 am/pm - 3.00 am/pm
Northern hemisphere (90N,0E, 18N,0W; DET: 1x1, EPS: 1.5x1.5 degrees)			
H8D - GRIB DET 00&12 UTC	21	2*66	10.00 pm/am - 1.00 am/pm
H8E - GRIB EPS 00&12 UTC	21	2*8	1.00 am/pm - 3.00 am/pm
Weather parameter BUFR files:			
H3A - BUFR DET WORLD	1	2*5	1.00 am/pm - 2.00 am/pm
H5A - BUFR DET HUNGARY	1	2*1	"
H5B - BUFR EPS HUNGARY	1	2*1	"
H6B - BUFR EPS WORLD	1	2*1	"
DATA DOWNLOADING FROM MARS			
Monthly Forecast for Hungary	4		
Seasonal Forecast for Hungary	5		

10 Report on the 16th meeting of Member State Computing Representatives, 21-22 April 2004, Hungary

HUNGARY

HUNGARY



IRELAND

IRELAND

Paul Halton – Met Éireann

Developments in 2003 (1)

During 2003

- **C4I Special Project established at Met Éireann**
  - A total of 6 staff joined the C4I project in the past year
  - HIRLAM model converted to run as a Regional Climate Model (RCM)
  - Funding streams have been established
  - All 6 staff registered as users of ECMWF computing facilities
  - Project currently utilises ECMWF HPCF platform

July 2003:

- **34 new Fujitsu-Siemens PCs purchased from Calyx**
  - Win-XP and Linux compatible

Nov 2003

- **MSG system**
  - Installed with EUMETcast DVB
  - Site Acceptance Tests completed with VCS engineer
  - VCS applications stable after main server memory was doubled

27/9/04 Met Éireann, Dublin. Ireland 2

Developments in 2003 (2)

NWP Systems:

- In April 2003, IBM engineer, Peter Mayes upgraded AIX, Loadleveller, PSSP and other system software on the RS/6000 SP server, SWIFT.
- During 2003, hardware failures occurred on a number of nodes on the SWIFT system and these were repaired by IBM under the terms of the maintenance contract.
- In one or two instances the operational HIRLAM cycle was interrupted
- It was necessary to utilise the output data from the backup NWP server which runs the NWP suite at a coarse resolution on a Dell dual processor Precision 530 MT, Xeon 2GHz PC running Red Hat Linux version 7.1.

27/9/04 Met Éireann, Dublin. Ireland 3

Developments in 2003 (3)

In Dec 2003, a contract was awarded to Dell for the supply of a seven-node Linux Cluster comprising…

- 1 x 42U Rack 4210 Base containing a Dell PowerEdge 1750 – AC
- 1 x master node: 2 x Xeon 2.8Ghz/512k 533Mhz FSB, 4Gbytes ECC DDR memory
- 6 x slave nodes: 2 x Xeon 3.2Ghz/1MB 533Mhz FSB, 2Gbytes ECC DDR memory per node with
- Scali interconnect and Dolphin cards.
- 1 x tape backup unit - PowerVault Tape System PV112T VS80 Rack Base 1U Single 40/80GB
- RedHat ES 3.0 on master node, WS 3.0 on compute nodes
- PGI Cluster development kit & Intel Fortran compiler

27/9/04 Met Éireann, Dublin. Ireland 4

IRELAND IRELAND

Developments in 2003 (4)

Projects currently under development include:

- **TUCSON – 9/25 x AWS Developed & Installed**
- **CAIRDE-2000 – Project completed in 2003**
  - **Link to Irish Aviation Authority in Ballycasey (near Shannon)**
  - **FDPS Data dissemination**
- **MSG / EARS – VCS awarded contract in 2002**
- **VAX-Cluster Replacement – Self Briefing for Pilots**
- **C4I Project – Regional Climate Model**
  - **Climate Change research project includes link to CosmoGrid Ireland**
  - **Gateway to GRID Computing resources still being researched**
  - **ERA-40 Data from ECMWF**
  - **Experiments will continue to be submitted to run on ECMWF HPCF system**

27/9/04 Met Éireann, Dublin. Ireland 5

Developments in 2003 (5)

- **Desktop Systems**:
  - Busy year regarding deployment, maintenance, support and protection.
  - A total of 315 PC-based systems are in operational use and about half are used as servers, workstations or are attached to instruments
  - From April 2004, all desktop PCs are automatically updated daily with the latest anti-virus protection software.
  - The GASP software asset management utility is used to maintain a hardware and software assets database.
  - A University student was employed for 6 months to:
    - Deploy 34 new Win-XP PCs
    - Compile a report on Desktop PC deployment Strategy.
  - After the IT asset inventory was completed an independent assessment of the internal audit procedures was made on behalf of Microsoft.
  - The auditor reported to Microsoft that the software asset management arrangements at Met Éireann were very detailed and exemplary.
  - With an ever increasing amount of SPAM e-mail and computer viruses in circulation on the Internet the Sophos anti-virus protection is automatically updated very three hours on the servers.

27/9/04 Met Éireann, Dublin. Ireland 6

Developments in 2003 (6)

DMZ on HQ Network:

- A new **D**e-**M**ilitarised **Z**one was set up in Q4, 2003 to facilitate:
  - two new iTouch servers,
  - a new Heritage Library server – to enable outstations to access catalogues
  - a new ECACCESS server.

ECACCESS Server operational from 20 Jan 2004

- ECACCESS server was successfully set up in the new DMZ area
- Much time was spent attempting to set up and test the ECACCESS server using RedHat version 9.0 and the latest version of Java.
- However, the gateway application only worked with RedHat Linux 7.2 or SuSE Linux 8.2 and with a specific old version of Java.
- Many thanks to Laurent Gougeon and colleagues for their patience and valuable assistance.
- Specific users (e.g. C4I Project team) were provided with instructions on how to use the ECACCESS server to connect to ECMWF over the Internet.
- All relevant scripts were updated on the ECMWF servers to replace the old ECCOPY commands with ECTRANS commands.

27/9/04 Met Éireann, Dublin. Ireland 7

Developments in 2003 (7)

RMDCN Link

- Equant upgraded the RMDCN bandwidth from 128kbps to 384kbps in Oct 2003!
- 00z data restored to dissemination schedule
- The HPCF computers at ECMWF will be utilised more in the future as the system architecture is similar to the IBM RS/6000 SP installed at Met Éireann
- No problems reported on the RMDCN link since it was upgraded

27/9/04 Met Éireann, Dublin. Ireland 8

Usage of ECMWF Facilities

Irish Usage of HPC Facilities at ECMWF

ECMWF Account Report for 20040419

```
Member state: Ireland
     allhpc 4 week report for period ending (MMDD): 0419
Node:  ie            SBUs:       0.0  total:   78389.9  alloc:  894000.0(8%)

Acct:  iebipw        SBUs:       0.0  total:       0.0  alloc:  107000.0(0%)
Acct:  iec4i         SBUs:       0.0  total:   43471.2  alloc:  100000.0(43%)
Acct:  iewind        SBUs:       0.0  total:   34918.7  alloc:  100000.0(34%)
```

27/9/04 Met Éireann, Dublin. Ireland 9

Future Plans at Met Éireann (1)

- Implement plans to improve forecast office efficiency – to include
  - Development of a Point Forecast Database
  - Automatic faxing system to send forecasts directly to customers
  - Procurement of a forecaster workstation system
  - Enhance TV Graphics facilities by deploying Borealis on SuSE Linux from WeatherOne systems
- Obtain ISO 9002 Accreditation for Aviation Services Division
- Replace Vax-4200 Cluster & re-develop existing applications to utilise RDBMS & Web-based technology.

27/9/04 Met Éireann, Dublin. Ireland 10

IRELAND IRELAND

Future Plans at Met Éireann (2)

- Download ECMWF 15 Years Re-Analysis data for Special Project, C4I.
- Continue development work on Linux Cluster
- Develop a strategy for the operational introduction of BUFR encoding and decoding of observation data transmitted and received on the RMDCN circuit.
- Prepare for the future replacement of T4-FAX products with alternatives for aviation users

27/9/04 Met Éireann, Dublin. Ireland 11

Use of ECMWF Facilities (1)

- Currently, ECMWF computing facilities are used as follows:
  - Data retrieval from the MARS archive
  - Experimental runs of the HIRLAM Model
  - Trajectory Model
  - Running MetView in batch mode
  - Boundary Condition data for HIRLAM
  - Training Courses

27/9/04 Met Éireann, Dublin. Ireland 12

Use of ECMWF facilities (2)

- **ECMWF Web Site:**
  - Users like the layout & features of the ECMWF Web site
  - The “Your Room” facility is popular with some users in the forecast offices and Research Division in particular.
- **Special Projects:**
  - Two special projects commenced early in 2003
    - C4I Project – (SPIEC4I) Based at Met Éireann HQ
    - EPS Verification – (SPIEUCC1) Based at Cork University
  - A new Special Project from MRCC, Cork, is proposed for later in 2004 – Its at Application form stage!!

27/9/04 Met Éireann, Dublin. Ireland 13

Use of ECMWF facilities (3)

- **RMDCN Link upgraded from 128kbps to 384kbps**
  - The HPCF computers at ECMWF will continue to be utilised more in the future
- **MetView & Magics:**
  - MetView & Magics are installed on SuSe Linux workstations at HQ – A new server will be used instead.
  - Usage permission granted to C4I project researcher based at UCD
- **ECaccess facility:**
  - Users have not reported any problems with new server!

27/9/04 Met Éireann, Dublin. Ireland 14

Future Use of ECMWF facilities

- **Dissemination Schedule:**
  - 00z run was restored after the Link was upgraded in October 2003.
- **GRIB Edition-2 and BUFR:**
  - Adequate notice of future ECMWF plans.
  - Access to GRIB Ed-2 decode software and sample test data for testing.
- **Training Courses:**
  - C4I Project members attended recent courses

27/9/04 Met Éireann, Dublin. Ireland 15

ECMWF User Comments (1)

Wind-Energy Special Project

- The Wind-Energy Special Project (EU-HONEYMOON) team based at UCC wishes to thank ECMWF for the use of the HPC facilities over the past year.
- At the end of the year Met Éireann allocated some its unused units to this Special Project that was vital for the development of SEPT, the Statistical Ensemble Prediction Tool that they are responsible for in the 5th Framework project, HONEYMOON.
- With the additional units they were able to produce an 8 month ensemble dataset of 50 members.
- They would also be happy to spend this year's excess resources on the HPCF system!

27/9/04 Met Éireann, Dublin. Ireland 16

IRELAND IRELAND

ECMWF User Comments (2)

Last year I ran a parallel Hirlam experiment [an extended integration with two versions of the model].

- My main problems were associated with Hirlam itself and most of my questions were to Hirlam people [i.e. G. Cats].
- One problem was **lack of scratch space on ecgate1** [the default allocation was not large enough to run Hirlam] and John Greenway increased it for me.
- I log into ECMWF from home via the Internet [using the SecurID card]. This works very well.
- Generally, my experience has been good.

27/9/04 Met Éireann, Dublin. Ireland 17

ECMWF User Comments (3)

- In general the C4I Project Team has been very pleased with the use of ECMWF computer resources, particularly with the turnaround of batch jobs on the IBM HPC.
- Retrieving data from experiments has been a bit of a bottleneck.
- One of the C4I project team, brought over to ECMWF 2 x 250GB external hard drives while attending a training course recently and User Support kindly downloaded a substantial volume of data onto the drives for him; it saved a potentially large load on the network.
- We would like to do this again, perhaps on a regular basis! What is the best way to achieve this?
- Also, would ECMWF facilitate having the hard drives posted to them?
- On ecgate1 the user space for $SCRATCH is rather small (~1GB) – this is an important issue when using "ecfsdir" to retrieve large tarred directories. Are there any plans to increase the size?

27/9/04 Met Éireann, Dublin. Ireland 18

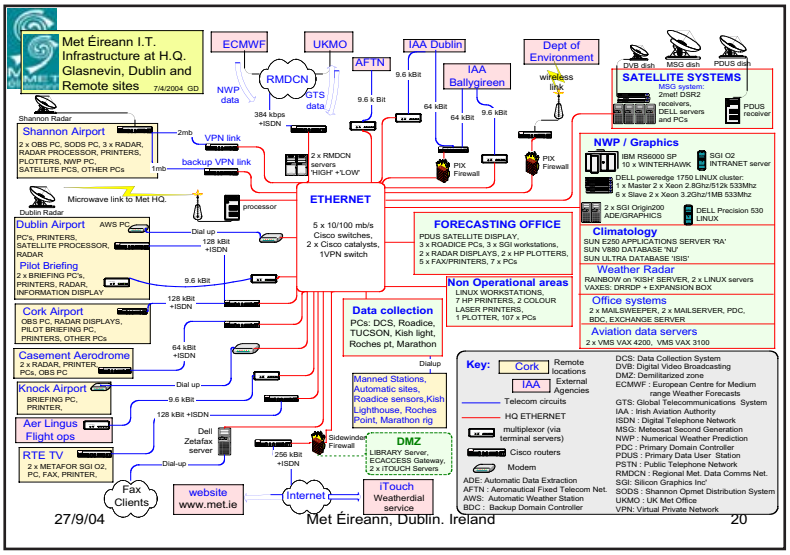
ECMWF User Support

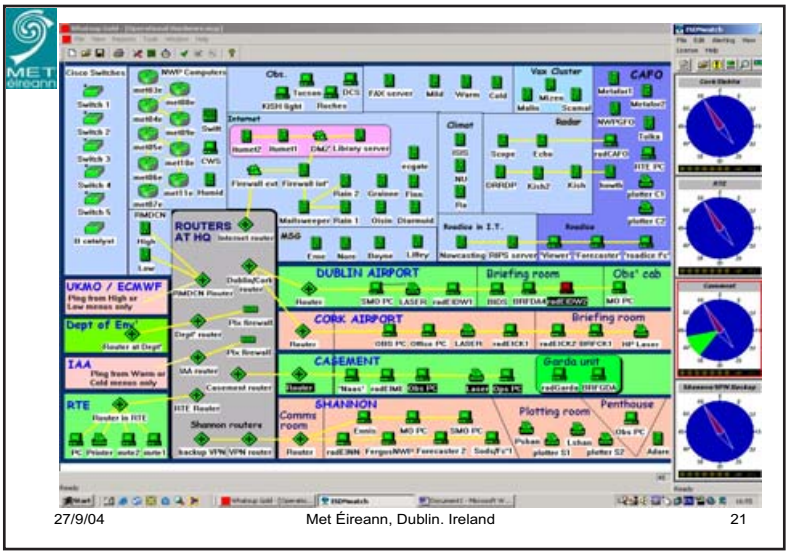
- As new staff joined the C4I project during the year they were each supplied with ECMWF SecurID cards.
- The C4I project was registered at the Centre as a Special Project from Ireland.
- Metview and Magics software licences were made available to one of the C4I project researchers located at University College Dublin for the duration of the Project.
- Norbert Kreitz, visited Met Éireann on 30 June & 1 July. He gave a very interesting talk on the collaboration between the two organisations over the years.
- Thanks to everyone in User Support and particularly a special thanks to Norbert in recognition of the excellent support he has given Irish users over the past 22 years.

27/9/04 Met Éireann, Dublin. Ireland 19

IRELAND

IRELAND

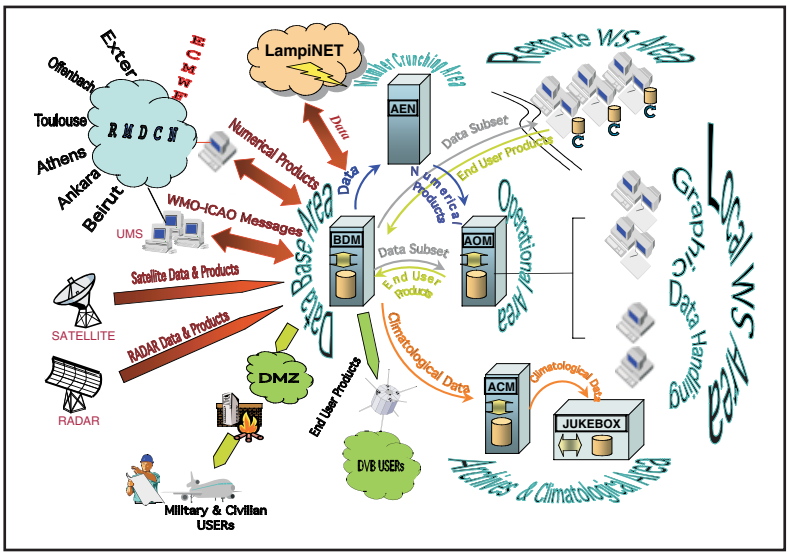




ITALY

ITALY

Giuseppe Tarantino – Italian Meteorological Service



Experience

More or less **115 users** using ECMWF services, most of them use INTERNET access:

- **University**
- **Regional Meteorological Services**
- **Research Agencies**
- **Environmental Agencies**
- **Armed Forces**
- **Environmental Hazard Department**

The main usage of ECMWF services is retrieval of MARS data associated with the decoding software to run either models or MAGICS and METVIEW applications.

ECMWF GRIB data

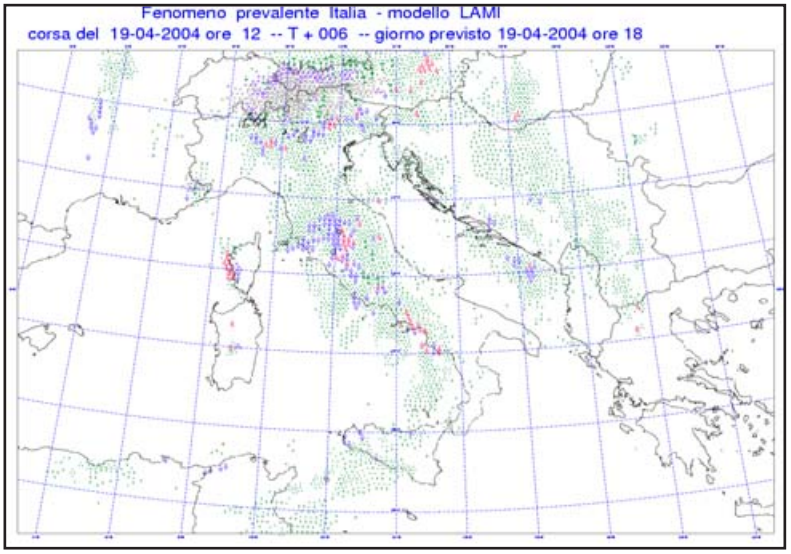
- Are routed in real/delayed time to Special Users for their operational duties (enviromental hazard, agriculture, pollution etc.)
- At the Operational Center are also used as:
  - support for the operational meteorological activities
  - boundary condition for the local models
  - input for post processing programs
  - input to produce information and maps useful for aeronautical duties

Projects (1/3)

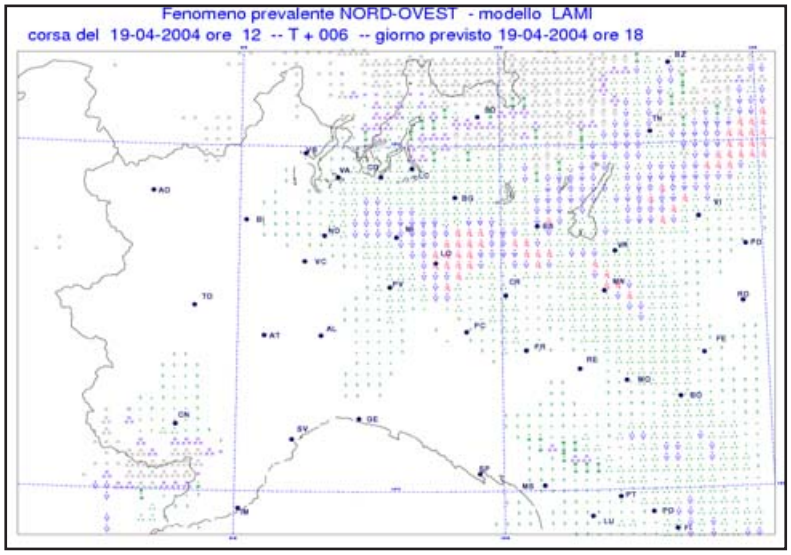
A test suite of the Lokal Model runs daily at ECMWF to verify the impact of Boundary Condition (originated from IFS) versus the operational version running with BC from GME. Moreover the model is tested by introducing additional parameters like cloud ice content. A selection of these products is archived on the Member State MARS area.

Projects (2/3)

An operational suite of the Lokal Model fields post-processing algorithm (AWI - Automatic Weather Interpretation) runs daily at ECMWF in order to get automatic weather interpretation in standard synoptical-like maps and data format (BUFR).
An objective study about the reliability of post-processing algorithm is in progress and we trust to obtain a full set of verification results by the end of this year.



ITALY **ITALY**



Projects (3/3)

An increased resolution (0.25° grid spacing) version of the hydrostatic HRM model is integrated up to +72h over the Euro-Atlantic domain.
The model runs on the HPCA platform using 14 processors in MPI mode. Boundary conditions are from IFS.
Initial conditions for the model are from the Italian Air Force Met Service 3D-Var assimilation system, which is also run on the ECMWF HPCA using 60 processors in MPI mode.

The special projects

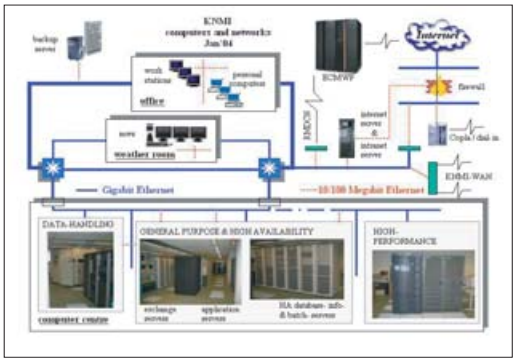
- EVALUATION OF THE PERFORMANCE OF THE ECMWF METEOROLOGICAL MODEL AT HIGH RESOLUTION
- NON LINEAR ASPECTS OF THE SYSTEMATIC ERROR OF THE ECMWF COUPLED MODEL
- LIMITATED AREA MODEL TARGETED ENSEMBLE PREDICTION SYSTEM (LAM-TEPS).

NETHERLANDS

NETHERLANDS

Hans de Vries – KNMI

Computer Infrastructure



Computer Infrastructure: highlights

- Sun Fire 15000 (68 CPUs)
- StorageTek PowderHorn 9310 tape silo (≤ 800 TB)
- Linux Workstations (SGI, Compaq, HP)
- Citrix servers for Windows applications in a UNIX environment
- 100 Mbit/s – 1 Gbit/s internal network

Computer Infrastructure: Main connections

Connection		Capacity [bit/s]
Internet		1 G
	Firewalls	4 x 100 M
RMDCN	Access line	384 k
	ECMWF	256/128 k
	Met Office (GTS)	64 k

NETHERLANDS

NETHERLANDS

Computer Infrastructure: developments in 2003

- New observations network
- New infrastructure for image data (HDF-5)
- Upgrade the mass storage system (30 – 800 TB)
- Upgrade and stabilization of the Citrix servers
- Upgrade RMDCN capacity
- Optimize firewalls

ECMWF MSCR-16: April 21-22, 2004, The Netherlands 5

Computer Infrastructure: developments in 2004

- New Internet servers
- New standard Linux distribution
- Start storage area network (data consolidation)
- Start migration to BUFR (NL observations)
- Upgrade firewalls?

ECMWF MSCR-16: April 21-22, 2004, The Netherlands 6

Users

Status on April 2

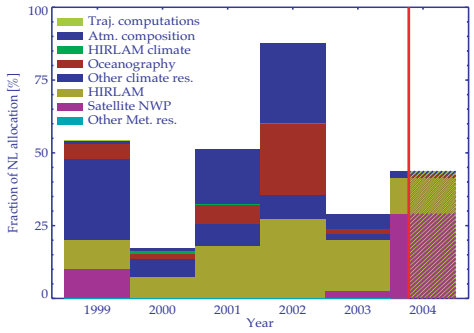
Total	83		
Last logged in since 1 March	52	Access to HPCF	59
Last logged in earlier in 2004	10	Access to Special Projects	11
Not logged in more than 1 year	8	Outside KNMI	13

ECMWF MSCR-16: April 21-22, 2004, The Netherlands 7

NETHERLANDS

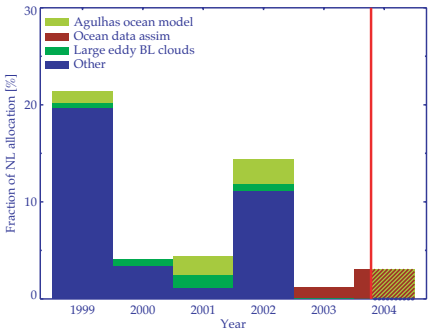
NETHERLANDS

Projects



ECMWF MSCR-16: April 21-22, 2004, The Netherlands 8

Special Projects



ECMWF MSCR-16: April 21-22, 2004, The Netherlands 9

Plans

Oceanography

- 20% of allocation

HIRLAM reconstruction

- Rerun latest HIRLAM for s 3 years
- 50% of allocation
- 15 TB data

ECMWF MSCR-16: April 21-22, 2004, The Netherlands 10

NETHERLANDS

NETHERLANDS

ECMWF software

EMOSlib
- Demand for GRIB/BUFR (de)coding software in C

SMS (not yet)

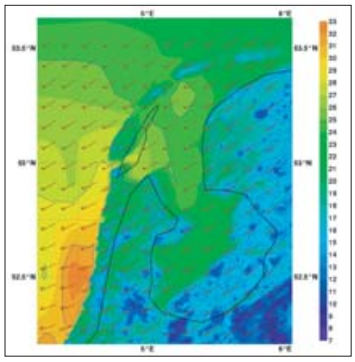
ECaccess gateway and tools
- Upgrade showed many improvements
- New way of X connections
- Database automatically converted
- Many ectrans transfers at the same time?
- Monitoring of the gateway?

prepIFS
- Xcdp over Internet big improvement
- Instability problems (Error 500) (Solved last Monday?)

Magics & Metview

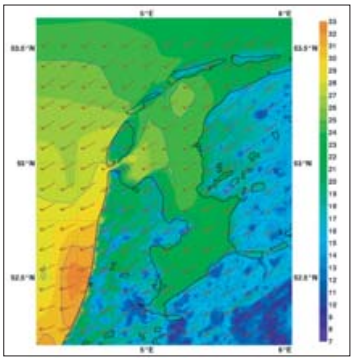
ECMWF MSCR-16: April 21-22, 2004, The Netherlands 11

MAGICS: old coastlines



ECMWF MSCR-16: April 21-22, 2004, The Netherlands 12

MAGICS: new coastlines



ECMWF MSCR-16: April 21-22, 2004, The Netherlands 13

NETHERLANDS

Linux

Distribution

- Now: Red Hat 7
- Red Hat Enterprise?
- Red Hat 9?
- SUSE?

Clusters

- High availability NFS server
- Blade technology
- Distributed computing

Comments

General

- The support from ECMWF is very much appreciated, e.g. User Support (John Greenaway), Call Desk (Petra did a great job in supplying many spare SecurID cards in time)
- More disk space required ($SCRATCH for HIRLAM, $HOME generally)

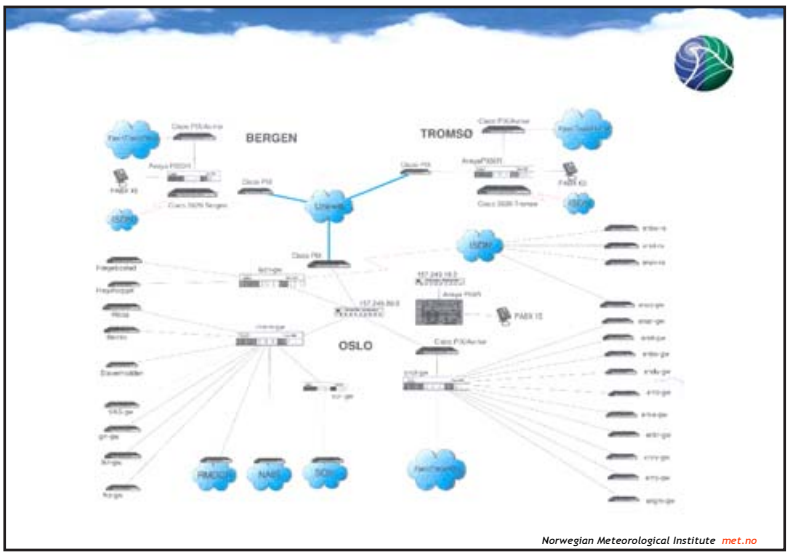
ECMWF web services

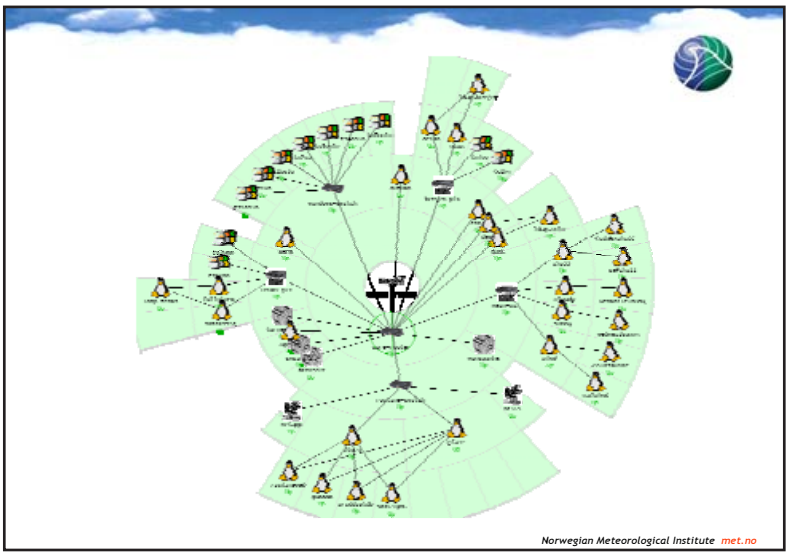
- Much very useful information (calendar, manuals)
- User (de)registration over the Web?
- How to find Computing Representatives?
- Access to restricted areas (TAC)?

NORWAY

NORWAY

Rebecca Rudsar – Norwegian Meteorological Institute, met.no





New since last time

- 2 Operational servers Dell PowerEdge 6650 with 4 x Intel XEON MP 1.90GHz, 4GB memory and 2 x 32 GB systemdisks (Raid). OS Debian GNU/Linux.
- Storage System Duplicated NAS-system from Network Appliance. 2 x F810 disks of 500GB. Full failover configuration.
- SMS now running under Linux SMS controls jobs running on remote servers.

Norwegian Meteorological Institute met.no

NORWAY NORWAY

New since last time (cont....

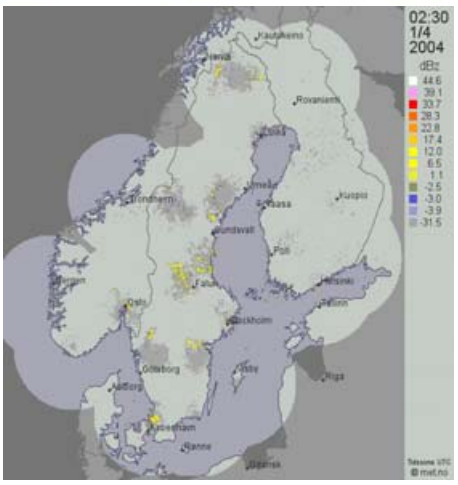
- Application servers Oracle IAS OS Redhat. eKlima is a portal for external users to extract data from the Climate Data Warehouse.
- Workstations. OS Redhat. Automatic installation via kickstart and configuration via cfengine.

Norwegian Meteorological Institute met.no

New since last time (cont....

- Dedicated Compaq servers. OS Debian. Satellite processing system, MEOS, will be operational 2.Q 2004.
- Dedicated Compaq server. OS Redhat. Nordic Radar system, NORDRAD, will be operational 2.Q 2004.
- New radar. To be installed on the island Rost in May.

Norwegian Meteorological Institute met.no



Norwegian Meteorological Institute met.no

New since last time (cont....

- Linux cluster. Accepted in March 2004.

Processers: 80 AMD Opteron 2.0 GHz assigned on 40 nodes.
Memory: 2 GB per node, total 80 GB.
Front-end: One dual node AMD Opteron 2.0 GHz.
Disk: approx. 740 GB, NFS-mounted on all nodes.
Interconnection: Myrinet for data, 100 Mbit Ethernet for administration.
OS: Redhat 9
Compiler: Portland Fortran and C
MPI & OpenMP: Scali
Queuesystem: OpenPBS

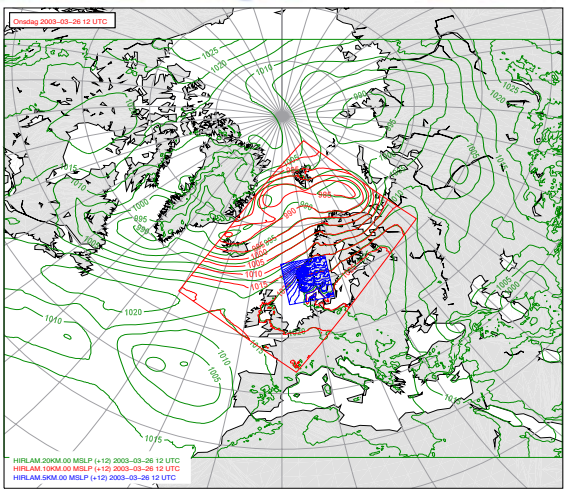
Norwegian Meteorological Institute met.no

Operational HIRLAM models

- HIRLAM 20 (0.2°, 468x378, 40 levels)
  3DVAR analysis, ECMWF boundaries
- HIRLAM 10 (0.1°, 248x341, 40 levels)
  Analysis from HIRLAM 20, ECMWF bnd.
- HIRLAM 5 (0.05°, 152x150, 40 levels)
  Nested in HIRLAM 10

All models updated Dec.2003 to version 6.2.0

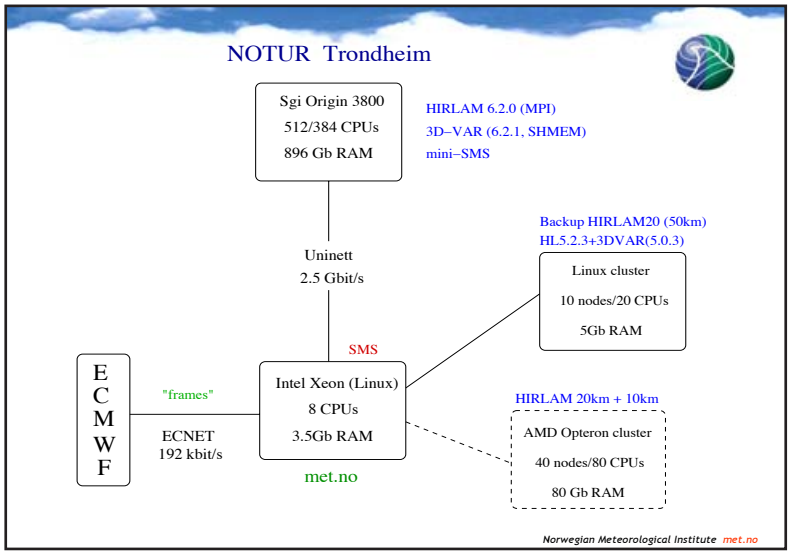
Norwegian Meteorological Institute met.no

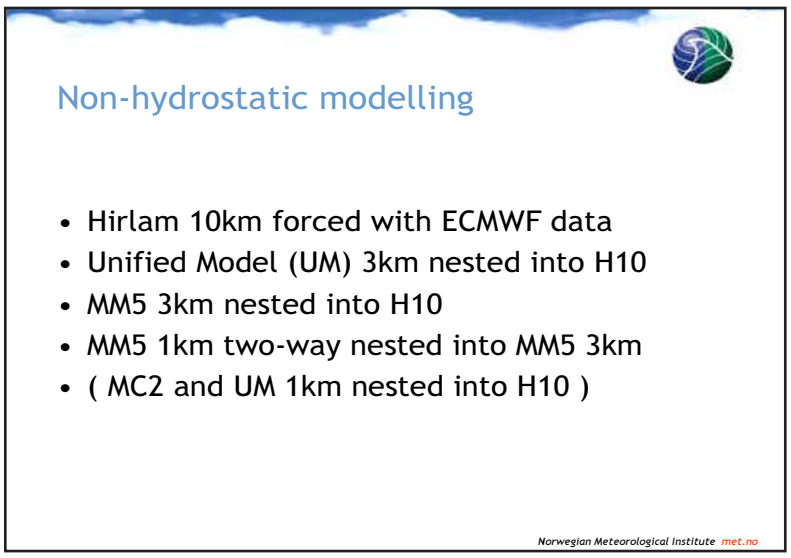


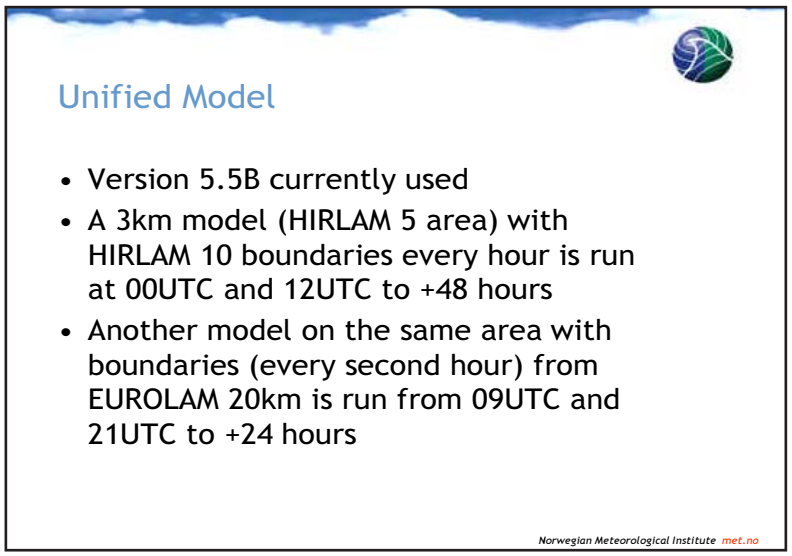
Norwegian Meteorological Institute met.no

NORWAY

NORWAY

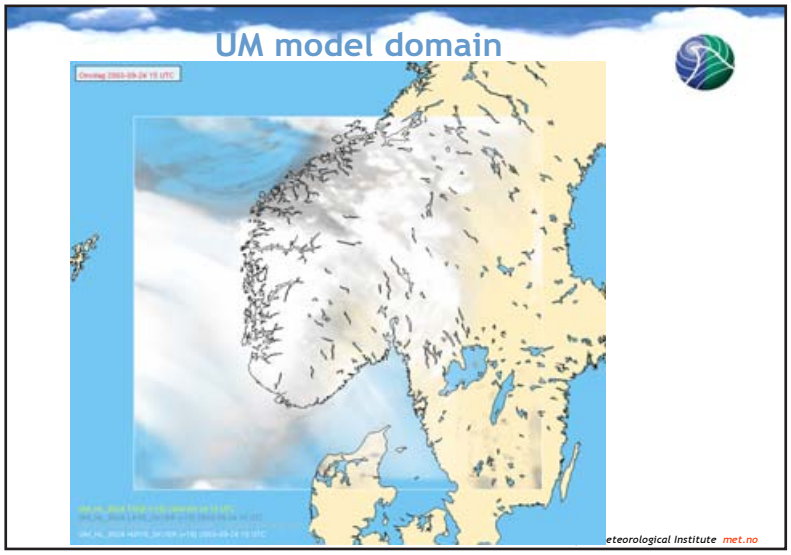


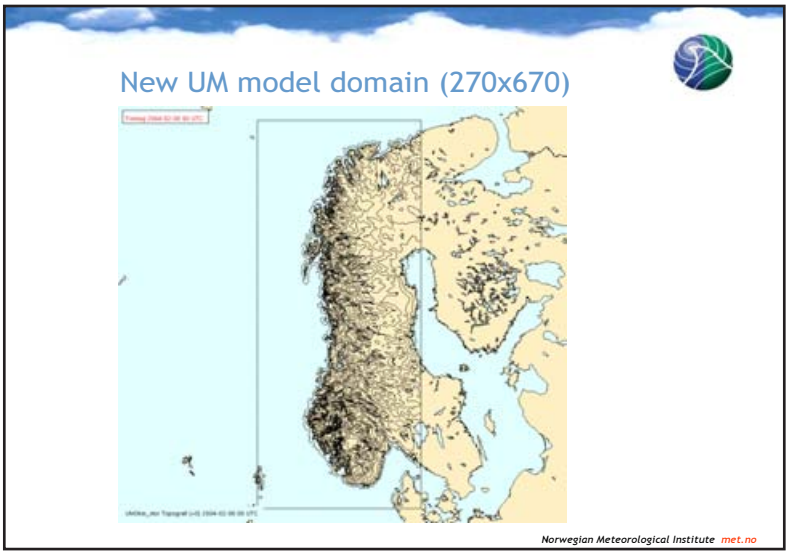




NORWAY

NORWAY





ECMWF Products

- via RMDCN

  DA, EF, Wave, BC1 and BC2 : 440 Mbyte

- via Internet

  DA and EF : 2260 Mbyte

Norwegian Meteorological Institute met.no

NORWAY NORWAY

ECMWF Projects

- Ozone as a climate gas.
- REGCLIM: Regional Climate Modelling.
- HIRLAM project.
- Targeted ensembles providing boundary values for limited area models.

Norwegian Meteorological Institute met.no

Feedback from the Users

- Total no. of users : 31
  operations = 2, met.no users = 23, external users = 6
- Total usage in 2003 : 30%
- In response to a request for feedback for this talk 11 users replied.
  7 use MARS, 2 use PrepIfs, 3 had no immediate plans for using ECMWF.

Norwegian Meteorological Institute met.no

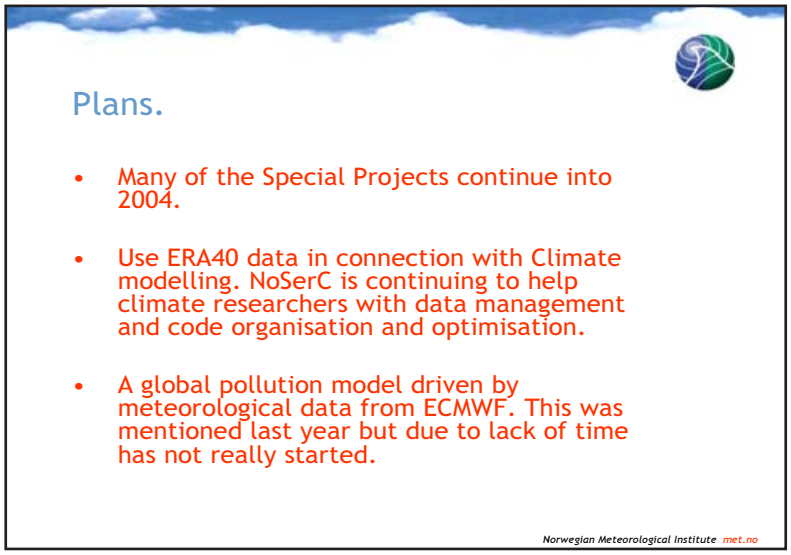
Comments from the users.

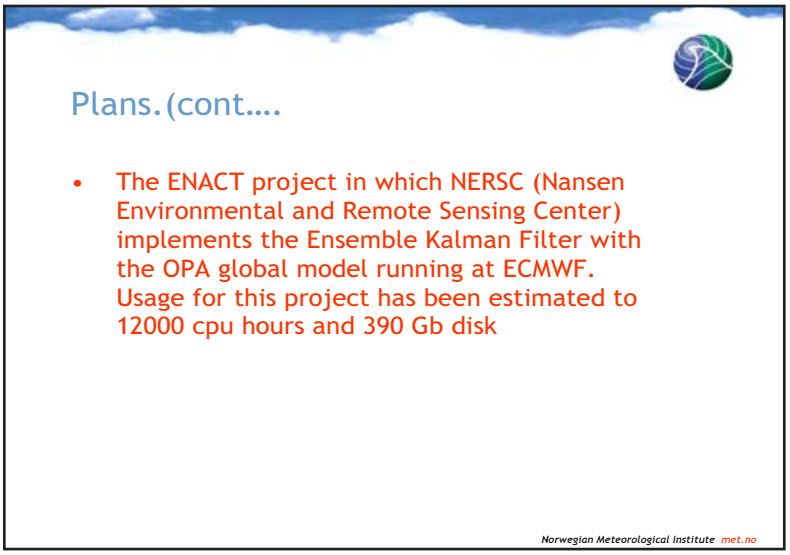
- "I have had problems with the change from Fujitsu to IBM because some of the IFS cycles were discontinued - in particular CY24R3."
- "As soon as ECMWF makes their products available in netCDF I will start to use them. Until then I will continue to use NCEP. Also NCEP products are available to everyone in the scientific milieu which means that other researchers can easily check the results of my research."
- Most users have no complaints and are very satisfied.

Norwegian Meteorological Institute met.no

NORWAY

NORWAY





ROMANIA

ROMANIA

Elena Toma – National Institute of Meteorology & Hydrology, Romania

National Institute of Meteorology & Hydrology, Romania

Computing facilities for operational computation of LAMs, weather forecast, research and development work.

- SUN servers
  - 1 E4500 (8 procs, 8 Gbytes memory)
    - used for running ALADIN limited-area model
  - 2 E3500 (4 procs, 2 Gbytes memory)
    - 1 used for running MM5 limited-area model
  - 4 Ultra 60 (2 procs, 1 Gbyte memory)
  - Blade 1000 (1 proc, 1 Gbyte memory)

Report on 16th Computing Representatives' Meeting, 21 - 22 April 2004

National Institute of Meteorology & Hydrology, Romania

- DEC workstations
  - 1 ALPHA 500
  - 1 ALPHA 250
    - used mainly for pre- and post-processing of NWP products
- HP servers
  - 1 Proliant (2 procs, 1Gbyte memory)
    - used for climatological database
  - 1 Proliant (1 proc, 500 Mbytes memory)
    - NOVELL server

Report on 16th Computing Representatives' Meeting, 21 - 22 April 2004

National Institute of Meteorology & Hydrology, Romania

- Linux servers
  - 3 (email, web, DNS respectively)
- Office PCs (> 300)
  - mainly under Windows 2000/NT4/XP
  - some using Linux (RedHat and Mandrake)

Report on 16th Computing Representatives' Meeting, 21 - 22 April 2004

ROMANIA ROMANIA

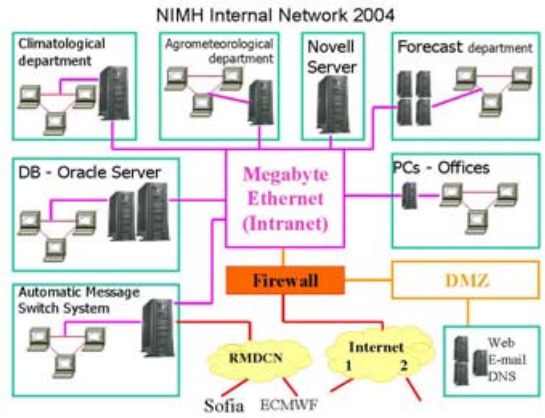
National Institute of Meteorology & Hydrology, Romania

Network environment

- Local Area Network
  - based on CISCO equipment (catalyst switchs, series 4000 and 3000) which provides 100 Mbits/s full duplex.
- Wide Area Network
  - connection between Regional Forecast Centers (RFC) and Central Operational Facilities (COF)
  - INTERNET (connection protected by firewall)

Report on 16th Computing Representatives' Meeting, 21 - 22 April 2004

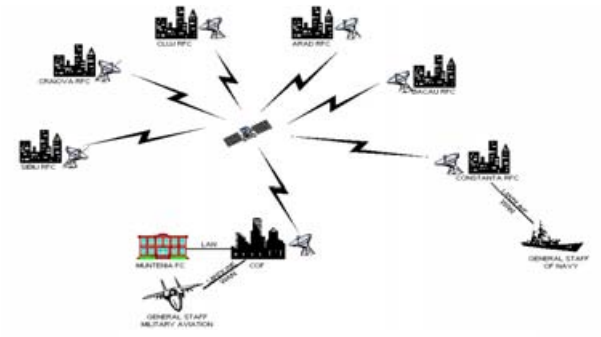
National Institute of Meteorology & Hydrology, Romania



Report on 16th Computing Representatives' Meeting, 21 - 22 April 2004

National Institute of Meteorology & Hydrology, Romania

• NIMH Wide Area Network - connection between RFC and COF



Report on 16th Computing Representatives' Meeting, 21 - 22 April 2004

ROMANIA

ROMANIA

National Institute of Meteorology & Hydrology, Romania

Plans for using ECMWF services

- Transferring deterministic model output (twice per day) and ensemble prediction results.
- Use of ERA-40 products
- using of METVIEW
- installing SMS

Report on 16th Computing Representatives' Meeting, 21 - 22 April 2004

SERBIA MONTENEGRO

SERBIA MONTENEGRO

Vladimir Dimitrijevic – Republic Hydrometeorological Service of Serbia

Computer resources for data/products receiving, processing and distribution

- Message Switching System (MSS) servers : Vax 3600 & Vax 400/200, Operating System (OS) VMS 5.5-2
- MicroVax II, OS VMS 5.5-2
- Work station Compaque Pentium 2+receiver/decoder + Satellite antenna, OS Windows NT4
- Local FTP server and TCP/IP: AMD Athlon(tm), OS Linux SuSe 7.1
- Domain Name Server, Proxy and Web server: IBM Pentium 4, OS FreeBSD 4.8
- Mail server, Firewall and Routing LAN: PC Pentium 3, OS Linux SuSe 7.3
- Dial-up and FTP Server: PC Pentium 2, OS Linux SuSe 7.3

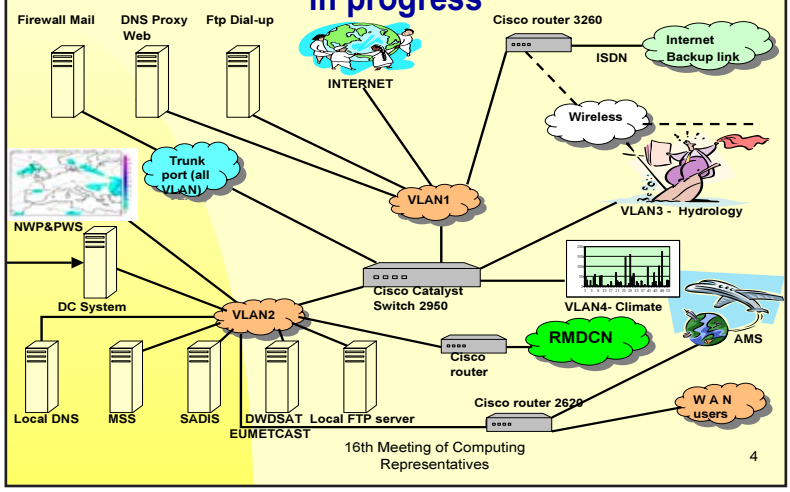
16th Meeting of Computing Representatives 2

Numerical Weather Prediction and Public Weather Service - Computer resources

Local Area Network	Jobs
Dual Pentium CPU2x1.0GHz	Eta model (DWD LBC) Resolution 52km/ 120 hours in advance
Pentium III CPU 600MHz	Eta model (AVN LBC) Resolution 52km/ 48 hours in advance
Sgi Indigo2 , Sgi O2 x 3, Sgi Indy, Sgi 550	Postprocessing, NCAR, GraDS, Archive,
BEOWULF cluster 5x4 CPU 1.4MHz	Ftp for LBC and Nonhydrostatic Eta model (DWD LBC) Resolution 18km/ 120 hours in advance.Postprocessing
IBM Xeon 2GHz x 2 (cluster)	Ftp of ECMWF products.Eta model (ECMWF LBC) Resolution 35km/ 72 hours in advance.Magics and MetView
Pentium III and Pentium IV (600MHz – 1.2GHz) x 20	Windows (98, 2000, XP) for applications, documents, print..., Public weather service
HP 2.4GHz x2	Linux SuSe 8.x for Back up and Research

16th Meeting of Computing Representatives 3

Network diagram of the RHMS of Serbia-Project in progress



16th Meeting of Computing Representatives 4

SERBIA MONTENEGRO

SERBIA MONTENEGRO

ECMWF products in operational use

- Products from deterministic forecast in GRIB based on 00Z and 12Z via internet
- Boundary conditions for limited area Eta model based on 00Z and 12Z via internet
- ECMWF software MetView, MAGICS, SMS
- MARS files on request
- Web available daily forecast including EPS

Data type	No.of products	size
SZD (BC)	70	2.6Mb
S1D (deterministic)	513	20.1Mb
S2D (global)	34	3.2Mb

16th Meeting of Computing Representatives

5

Future Plans for NWP

COMPUTER RESOURSE

- **BEOWULF CLUSTER 6x6, CPU 3GHz per node, Linux SuSe 8.x**
- **Work Station x12, CPU 3GHz, Linux SuSe 8.x**

NUMERICAL MODELS

- **Eta Hydro/Nonhydrostatic version with finer resolution**
- **Extended use of ECMWF products (Deterministic forecast, EPS, Seasonal forecast, Extreme Forecast Index…)**

16th Meeting of Computing Representatives

6

T. Lorenzen, noting that the RHSS received the ECMWF boundary conditions they needed to run ETA via the Internet, enquired the reliability and stability of the service.

M. Dell'Acqua replied that ECMWF had been maintaining statistics and, on average, there had been one outage per month per country. The re-establishment of the link after an outage was often slow, as so many companies are involved and no one wants to accept responsibility.

SLOVENIA

SLOVENIA

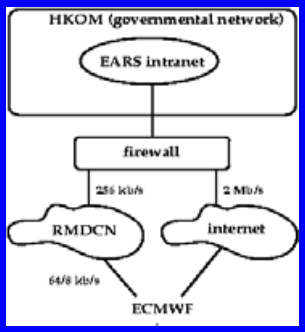
Miha Razinger – Environmental Agency of Republic of Slovenia (EARS)

Computers at EARS

- Linux cluster for Aladin/SI (Intel, RedHat, SCore)
- Old cluster (DEC Alpha, RedHat)
- Database server (Intel, RedHat)
- Workservers (Intel, Alpha, RedHat)
- Desktops (Intel, WinXP, Fedora)
- Connections: nfs, ssh, ftp, smb

Network at EARS

- EARS intranet
- HKOM network (strong policy)
- Internet & RMDCN lines to Reading



Use of ECMWF Services

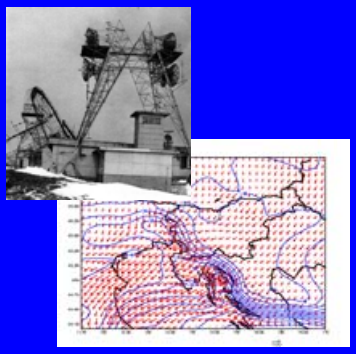
- No HPC facilities
- Daily transfers of deterministic model, EPS, multianalysis data ...
- SMS operational
- Web (Epsgrams,..)
- Occasional MARS retrievals

SLOVENIA

SLOVENIA

Future Plans with ECMWF Services

- Regional reanalysis of ERA-40 data with LAM Aladin/SI (4 TB)
- Preliminary tests promising



SPAIN SPAIN

Eduardo Monreal – Instituto Nacional de Meteorolgia

Instituto Nacional de Meteorología - Spain

1. Main computers

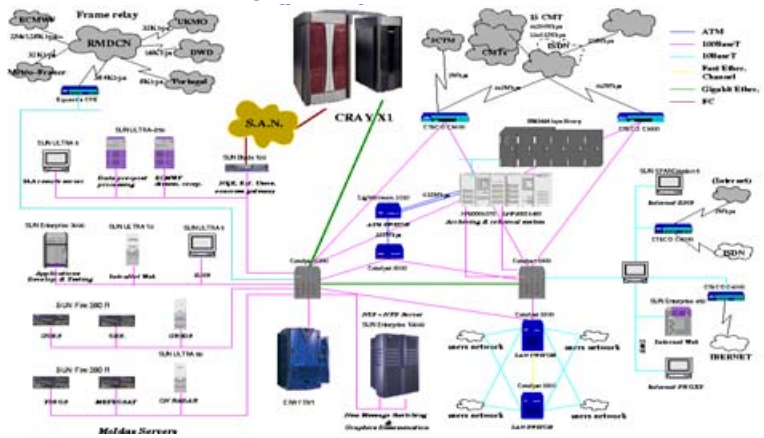
Major changes since last meeting:

- **The move to a new Computer Hall completed**
- **New high performance system: the CRAY X1**
- **A Storage Area Network implemented**

INM -Spain ECMWF 16th MS Computing Representatives' Meeting, April 21-22 2004

Instituto Nacional de Meteorología - Spain

Computer and Network



INM -Spain ECMWF 16th MS Computing Representatives' Meeting, April 21-22 2004

Instituto Nacional de Meteorología - Spain

New High Performance System. The CRAY X1

Initial system installed last August:

- **100 times the C94A**
- **10 computing nodes + 1 support node**
- **Each node has 4 multi-stream processors (MSP) and 16 Gbytes of high bandwidth (20.5 Gbytes/s) shared memory**
- **12.8 Gflops peak performance per MSP**
- **2 Mbytes of memory cache per MSP**
- **51.2 Gbyte/s full duplex 2D torus between nodes. Cache coherency globally addressable**

INM -Spain ECMWF 16th MS Computing Representatives' Meeting, April 21-22 2004

SPAIN SPAIN

Instituto Nacional de Meteorología - Spain

CRAY X1. Initial system specification (cont):

- **Distributed I/O: 4 SPC (1.2Gbytes/s) per node. On our configuration only support node is handling I/O**
- **Gigabit Ethernet connection through CNS (a Dell PowerEdge *2650* running LINUX)**
- **1.8 Tbytes of direct attached disk space (2 x 2Gb/s FC arbitrated loop)**
- **Cross compiling on CPES (8 CPU SUN Fire V480)**
- **One single system image: O.S. runs on support node only**

INM -Spain ECMWF 16th MS Computing Representatives' Meeting, April 21-22 2004

Instituto Nacional de Meteorología - Spain

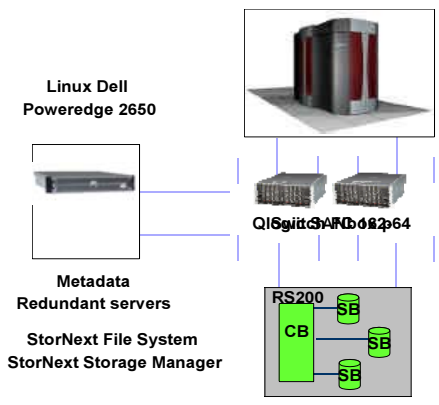
CRAY X1. Initial system specification (cont):

- **A Storage Area Network which consist of:**
  - FC switching equipment: 2 *Qlogic SANbox2-64 configured with 16 2Gb/s ports each*
  - 3 Tbytes of additional disk space
  - An ADIC scalar 100 robotic system (4 SCSI LTO-2 drives, 72 slots & 14.4 Tbytes of uncompressed data capacity)
  - ADIC's Stornext Management Suite (Stornext File System & Stornext Storage Manager -HSM-)
  - 2 Snornext FS metadata & HSM servers (Dell PowerEdge 2650 running Linux)

INM -Spain ECMWF 16th MS Computing Representatives' Meeting, April 21-22 2004

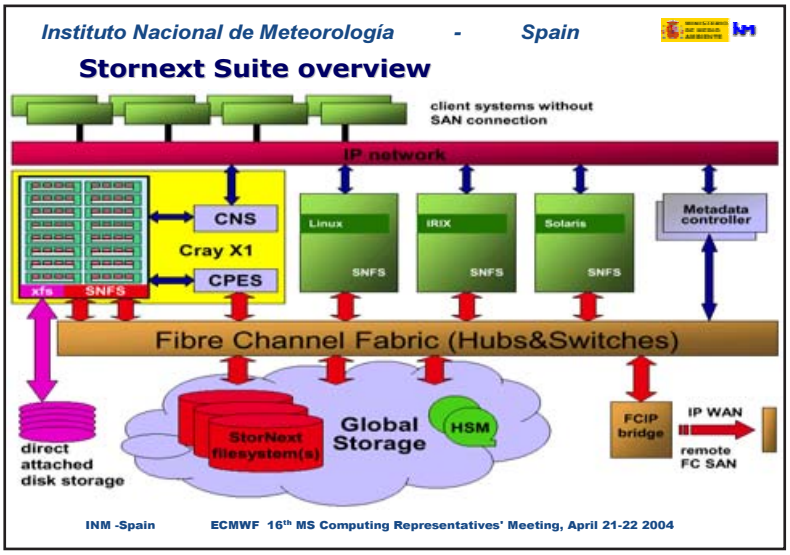
Instituto Nacional de Meteorología - Spain

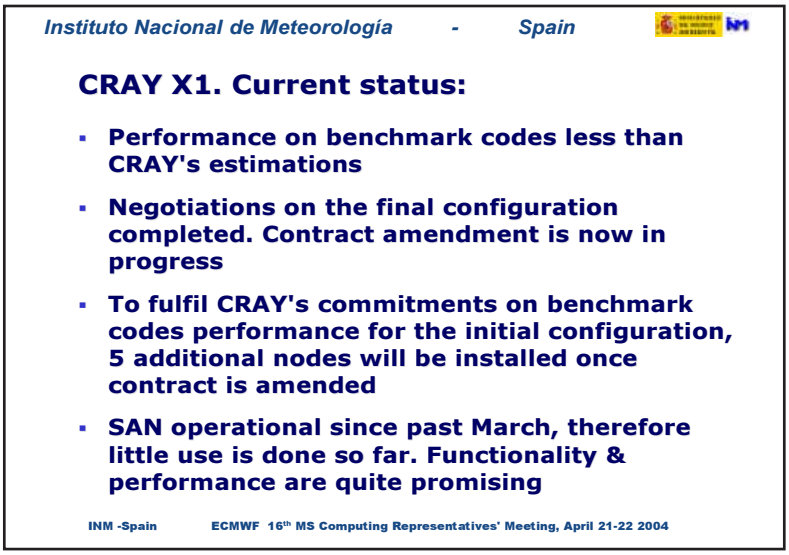
SAN layout

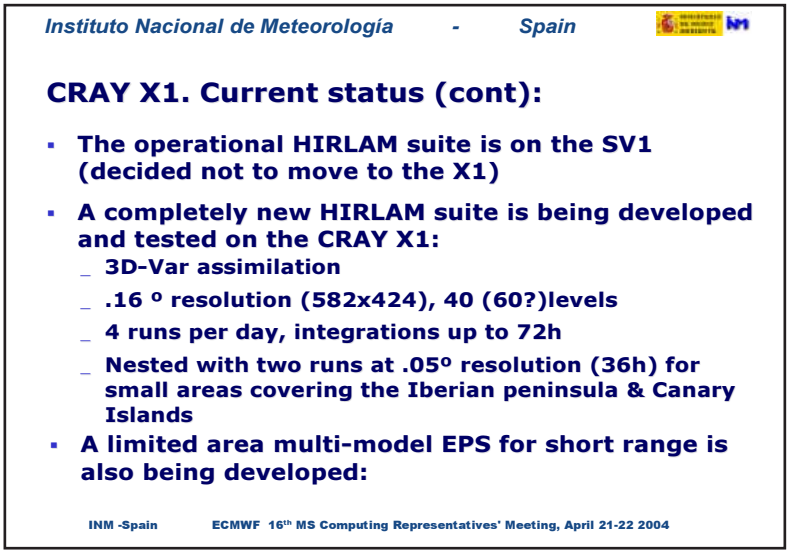


INM -Spain ECMWF 16th MS Computing Representatives' Meeting, April 21-22 2004

SPAIN SPAIN


Instituto Nacional de Meteorología - Spain
Stornext Suite overview
client systems without SAN connection
IP network
CNS
Cray X1
CPES
xfs
SNFS
Linux
IRIX
Solaris
SNFS
SNFS
SNFS
Metadata controller
Fibre Channel Fabric (Hubs&Switches)
direct attached disk storage
StorNext filesystem(s)
Global Storage
HSM
FCIP bridge
IP WAN
remote FC SAN
INM -Spain
ECMWF 16th MS Computing Representatives' Meeting, April 21-22 2004


Instituto Nacional de Meteorología - Spain
CRAY X1. Current status:
Performance on benchmark codes less than CRAY's estimations
Negotiations on the final configuration completed. Contract amendment is now in progress
To fulfil CRAY's commitments on benchmark codes performance for the initial configuration, 5 additional nodes will be installed once contract is amended
SAN operational since past March, therefore little use is done so far. Functionality & performance are quite promising
INM -Spain
ECMWF 16th MS Computing Representatives' Meeting, April 21-22 2004


Instituto Nacional de Meteorología - Spain
CRAY X1. Current status (cont):
The operational HIRLAM suite is on the SV1 (decided not to move to the X1)
A completely new HIRLAM suite is being developed and tested on the CRAY X1:
3D-Var assimilation
.16 º resolution (582x424), 40 (60?)levels
4 runs per day, integrations up to 72h
Nested with two runs at .05º resolution (36h) for small areas covering the Iberian peninsula & Canary Islands
A limited area multi-model EPS for short range is also being developed:
INM -Spain
ECMWF 16th MS Computing Representatives' Meeting, April 21-22 2004

SPAIN

SPAIN

Instituto Nacional de Meteorología - Spain

CRAY X1. Final configuration:

- **To be installed by 1Q 2005**
- **All 16 X1 nodes will be replaced by X1e nodes**
- **X1e nodes:**
  - **8 x 19.2 Gflops MSP**
  - **4 x 32 Gbytes & 12 x 16 Gbytes of shared memory**
  - **4 Mbytes cache memory per MSP**

INM -Spain ECMWF 16th MS Computing Representatives' Meeting, April 21-22 2004

Instituto Nacional de Meteorología - Spain

2. Connection to ECMWF

- **Following last RMDCN upgrade, access line is running at 384 Kbps**
- **Link to ECMWF has now a CIR of 256/128kbps (out/ in)**
- **Version 2.1.0 of Ecaccess gateway installed for both operational and users work on different platforms:**
  - On a Sun Blade 100 via the Internet for users, ectrans for the most part
  - On 2 Sun Ultra 250 servers via RMDCN for operational use (job submission, ectrans)
  - Java engine demands CPU intensively, at least on Solaris

INM -Spain ECMWF 16th MS Computing Representatives' Meeting, April 21-22 2004

Instituto Nacional de Meteorología - Spain

3. Experience using ECMWF computers (I)

- **Continues to be an upward trend in the number of registered users**
  - Currently 67
  - 58 last year
- **About 40 out of the 67 users are active**
- **Work done is for the most part MARS data retrievals, particularly access to ERA-40 dataset**
- **Metview used in batch mode to produce derived EPS products**
- **The new ecgate server hardly used so far**

INM -Spain ECMWF 16th MS Computing Representatives' Meeting, April 21-22 2004

SPAIN SPAIN

Instituto Nacional de Meteorología - Spain

3. Experience using ECMWF computers (II)

- **On 2003, 21 users accessed the High Performance Computers. They basically worked in the following areas:**
  - **HIRLAM model runs using the reference system**
  - **Trajectory computations**
  - **Studies on variability**
  - **Statistical downscaling of seasonal forecast outputs**

INM -Spain ECMWF 16th MS Computing Representatives' Meeting, April 21-22 2004

Instituto Nacional de Meteorología - Spain

3. Experience using ECMWF computers (III)

- **In 2003 the use of our HPCF allocation dropped to a 58%, distributed as follows:**
  - 90% HIRLAM runs,
  - 5% Seasonal Forecast downscaling,
  - 3% Studies on variability,
  - 2% Trajectory computation
- **In 2004 only used so far, less than a 6% of our allocation**

INM -Spain ECMWF 16th MS Computing Representatives' Meeting, April 21-22 2004

Instituto Nacional de Meteorología - Spain

3. Experience using ECMWF computers (IV)

- **Comments & queries from users (I got feedback from 12 users):**
  - Very satisfied, in general, of ECMWF computer services
  - Assistance & help from User Support, very much appreciated
  - Ecjls sometimes worthless, i.e. returns WAIT for jobs already executed (should be DONE). A remote qstat/llq would be preferable
  - Jobs submitted via ecjput result in STOP status often
  - A user ask for an sftp plugin into ecaccess gateway & he would like once an ssh session is established to have additional ssh term without authentication
  - Concerning LIBEMOS on LINUX, is there any plan to avoid the need of a FORTRAN 90 compiler?

INM -Spain ECMWF 16th MS Computing Representatives' Meeting, April 21-22 2004

SPAIN

SPAIN

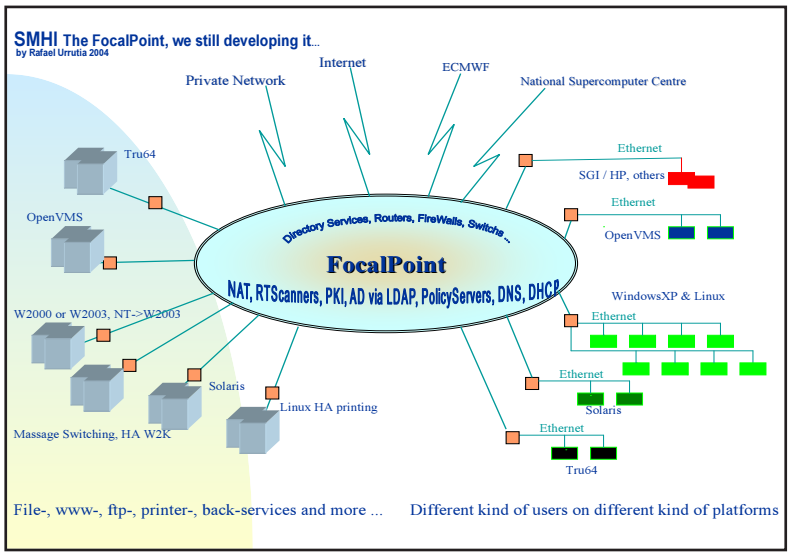
In reply to the question on LIBEMOS, F. Hofstadler explained that Ftn90 extensions were not used for BUFR and GRIBEX decoding, even in the latest version of LIBEMOS. Ftn90 was only used in interpolation.

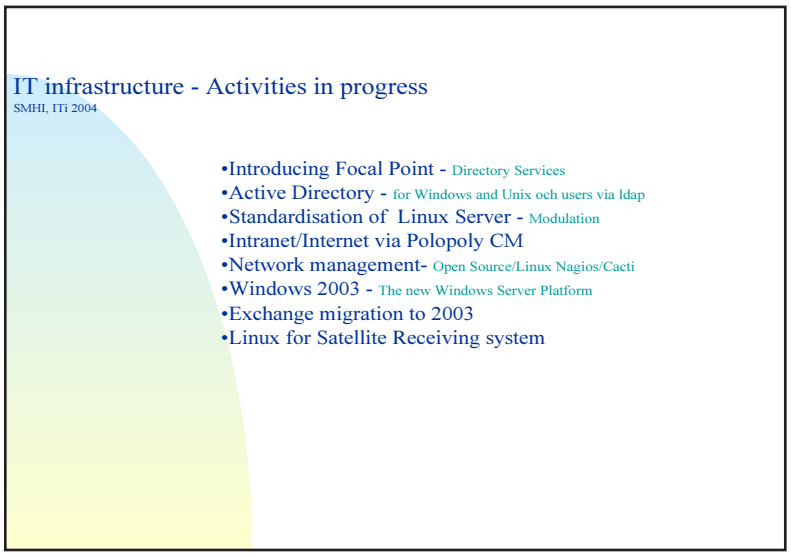
In reply to L. Gougeon's enquiry, E. Monreal clarified that the request for an sftp plug-in into the ECaccess gateway at ECMWF came from a university user who had no access to the INM gateway.

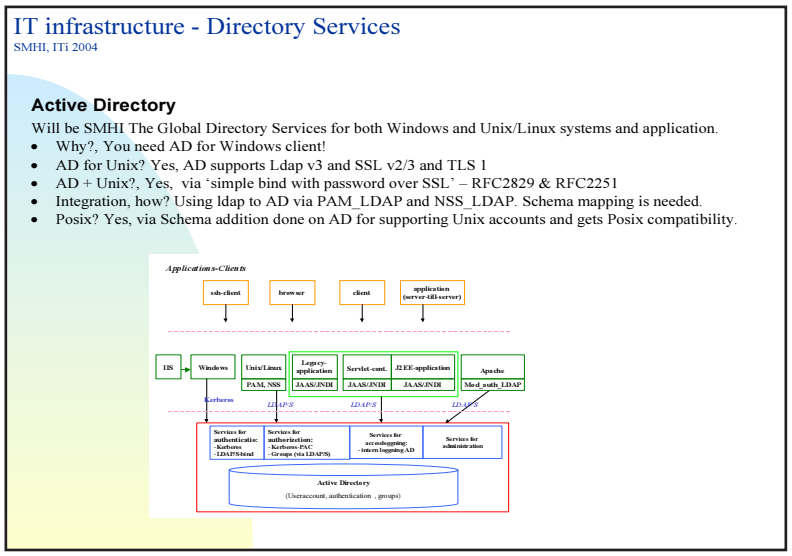
SWEDEN

SWEDEN

Rafael Urrutia – Swedish Meteorological and Hydrological Institute (SMHI)







SWITZERLAND SWITZERLAND

Peter Roth – MeteoSwiss

Switzerland
Peter Roth, MeteoSwiss, Zurich, April 2004

Actual Computer Environment

By the end of last year, we finished our project `Server, Storage & Client' (SSC). Now, we have a modern computer equipment. Actually, the main components are:

a) at Zurich

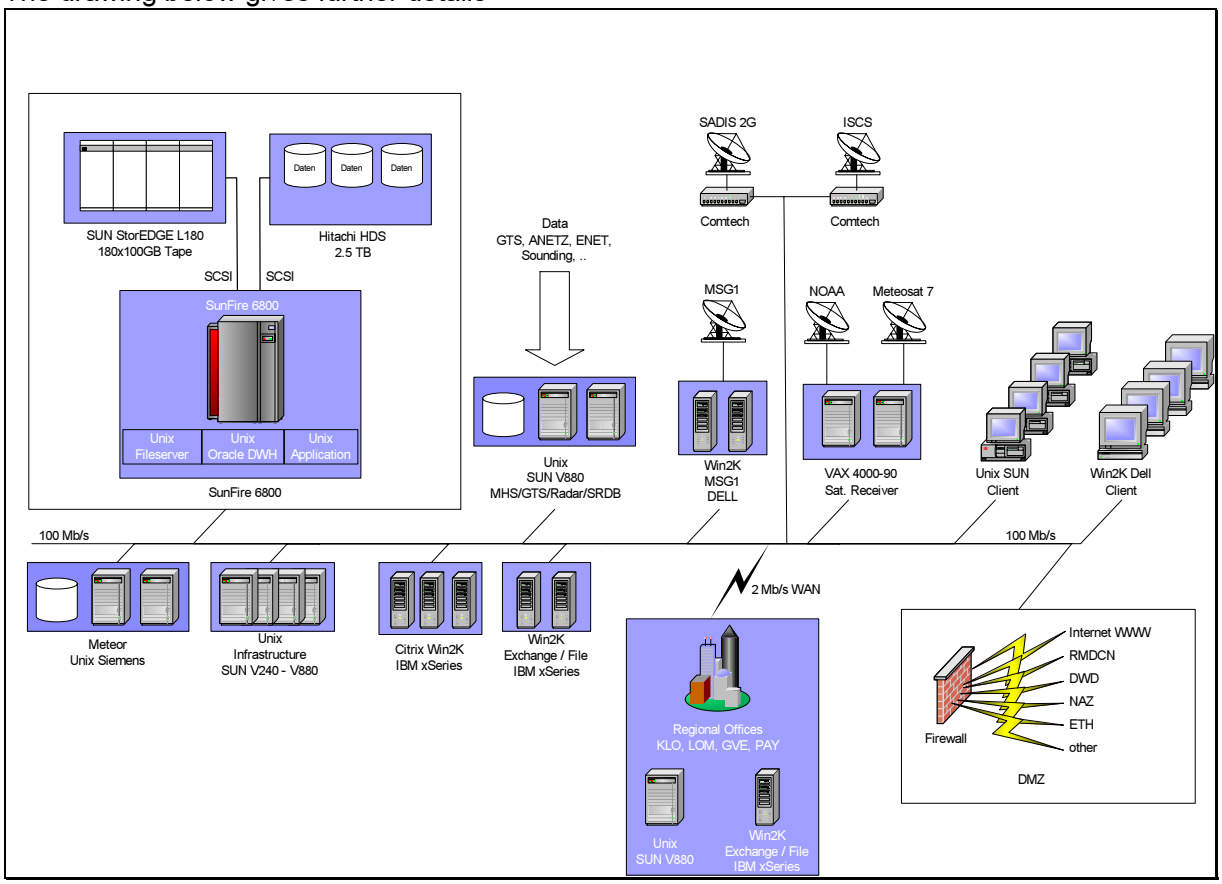
- The heart of the system is a SunFire 6800 with 4 domains running Solaris 8. One domain is used as fileserver controlling the NAS (Network Attached Storage) and the backup/recovery tapes. A second one is used as database server (Oracle) of our climatological database. Another domain is for producing meteorological and climatological products and the last one is scalable to increase the availability of the system.
- The servers for the message handling system (MHS) are SunFire V880 machines. On the same machines, we have integrated the radar servers and the data acquisition system.
- For receiving MSG-1 and DWDSAT data, we have installed a DVB receiver and a processing server running Windows 2000.
- No major changes were done in the DMZ.
- For office applications (‘MS-World') and mail services, we use servers from the IBM 345 Series, running an AD server, an Exchange server and a MetaFrame server. Users with a workstation (SUN / Solaris 8) use Outlook from the MetaFrame server.
- The old machines which still are operational are the ENET data acquisition system, the NOAA / Meteosat 7 system and the SADIS/ISCS system. The old mainframe 'Meteor' will be switched off in a few weeks.

b) at the Regional Offices (Zurich-Airport, Geneva, Locarno, Payerne)

The configuration is the same as at Zurich, but with smaller machines (SunFire V880 instead of SunFire 6800) and less storage capacity.

SMA ISM SMI MeteoSchweiz

The drawing below gives further details



Some figures

a) Equipment

- Unix Server: about 60
- Windows Server: about 20
- VMS Server: 4
- Unix Workstation: about 320
- PC (Windows): about 60

b) Network

- LAN: 100 Mb/s
- WAN: 2 Mb/s (we still have no backup-lines)
- ETH/CSCS: 10 Mb/s
- Internet: 5 Mb/s
- RMDCN:
  - ECMWF: 96 kb/s
  - DWD: 128 kb/s
  - MeteoFrance: 16 kb/s

Plans

There are plans to migrate to LINUX in about 2 years (first the servers, then the clients). Actually, we make studies how such a migration could proceed.

Experience using ECMWF Computer Services

MeteoSwiss makes use of:

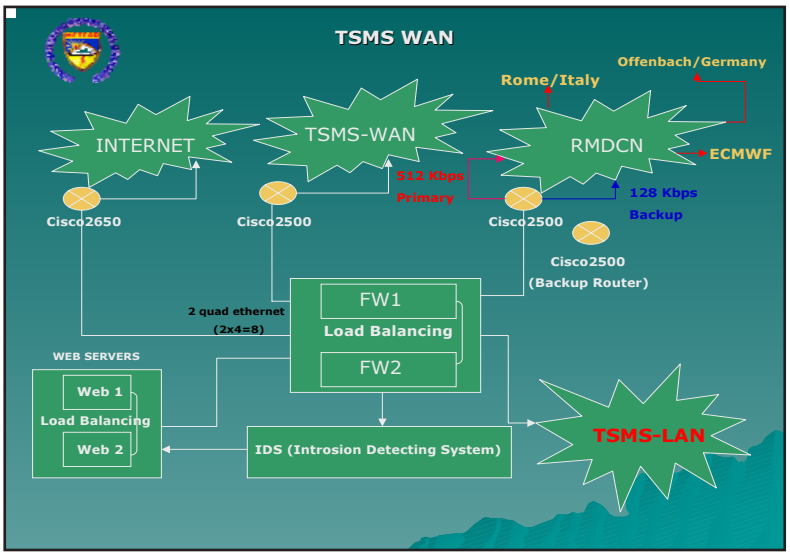
- the dissemination system (different data sets)
- MARS
- several services form ‘ecaccess'
- MAGICS applications running at ECMWF
- using MetView
- COSMO-LEPS calculations (producing products)
- Global EPS calculations (verifications and producing products)

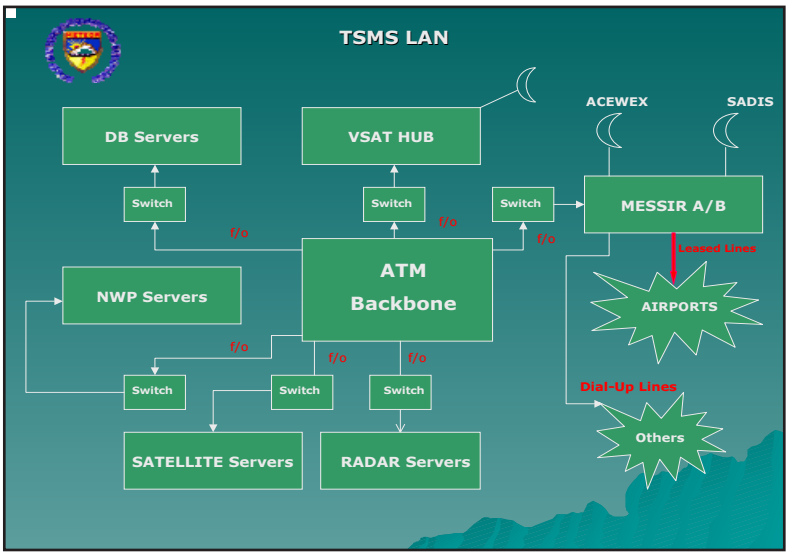
MeteoSwiss is a member of the ECMWF special project SPCOLEPS

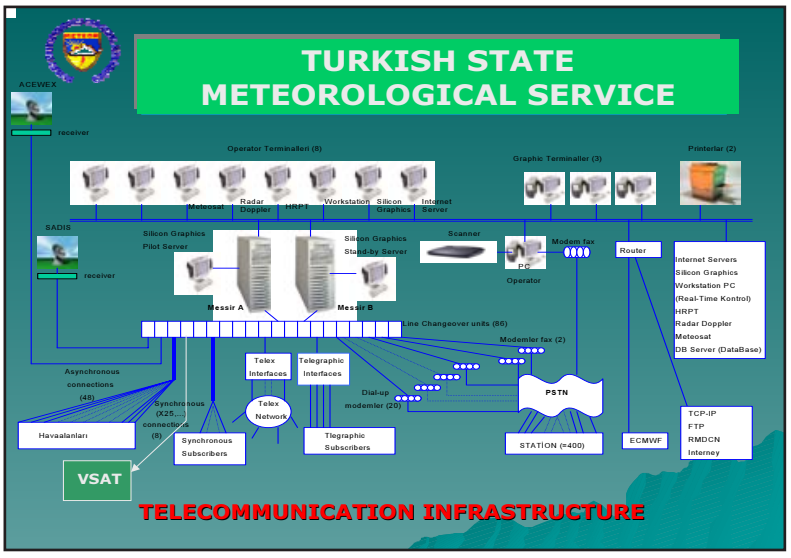
The users in Switzerland are very satisfied of the user support and the services from ECMWF.

TURKEY TURKEY

Bülent Yagci – Turkish State Meteorological Service (TSMS)







TURKEY

TURKEY

MSS COMPUTERS for TELECOMMUNICATION

Master and Slave 2 Slicon Graphics ORIGIN 200 Servers

- 2*MIPS RISC R-10000 64 bit CPU
- 2*9.1 GB HDD
- 128 MB MEMORY
- UNIX (IRIX 6.5)
- MESSIR software package

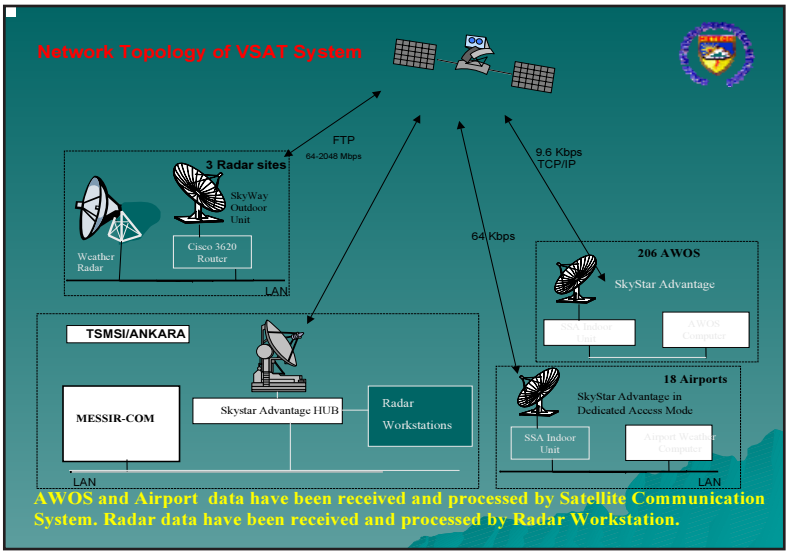
Data quality system based on code controlling is performed by both systems and software developed at TSMS.

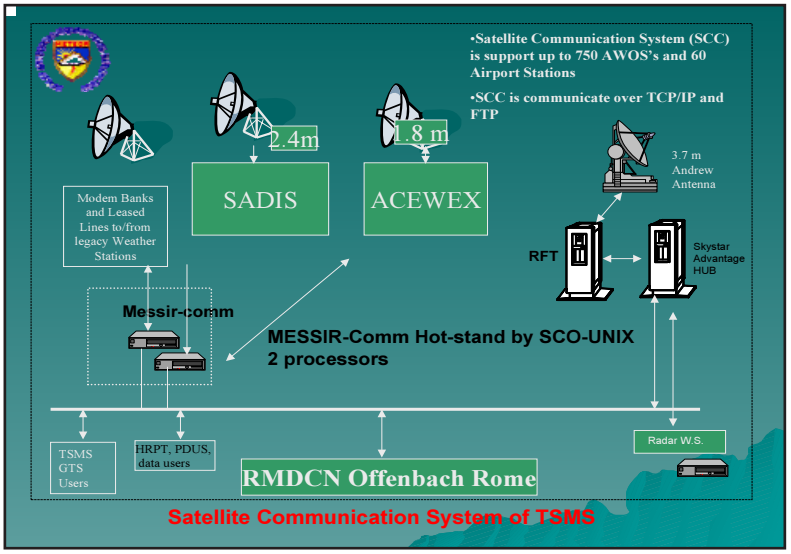
MSS COMPUTERS

FRONT-END PROCESSORS

- Intel Pentium II 350 MHZ
- 80 MB RAM, 4 GB HDD, CD-ROM drive*12
- 10/100 MB Ethernet
- DIGIBOARD card providing 80 asynchronous ports
- EICON Card providing 1 synchronous port
- MEGAPAC Switch providing 8 synchronous ports

An automated system called "MESSIR-COM" is used for data exchange over GTS and RMDCN.





TURKEY

TURKEY

DATABASE

Data Types:

1. Climatological Data

Meteorological parameters have been observed at local hours. 3 times a day or hourly.

Operating System: SCO Unix OpenServer 5.04

Database: Informix IDS 7.3 Relational Database Management System

Server: Compaq Proliant 6000
2*Pentium Pro 200 Mhz
512 MB RAM
56 GB Disk capacity

Data: Temperature, Precip. and humidity.

Data size on disk is about 7 Gb.

2. Upper-Air (Ravinsonde) data

Operating System: SCO Unix OpenServer 5.04

Database: Informix IDS 7.3 Relational Database Management System

Server: Compaq Proliant 6000
2*Pentium Pro 200 Mhz Processor
512 MB RAM
56 Gb Disk capacity

Data: 1971 – 1995: Significant Level height, temperature, dew point temperature, wind speed and direction.

1995 – Now: 400-900 Level Height, temperature, dew point temperature, humidity, wind speed and direction, inversion, tropopoz and max. Wind level data.

Data size on disk is about 3 Gb.

3. Synoptic Data

Operating System: SCO Unix Unixware 7.1

Database: Sybase ASE 11.9.2 Relational Database Management System

Server: 2*HP LH 6000
4*Pentium 3 700 Mhz
2 Gb RAM
360 Gb Disk capacity

Data:

Synoptic Data valid from 1980 for synoptic hours.
From the beginning of 2004, 200 AWOS Data have been collected and stored on Database Environment with 1 hour frequency.

Data size on disk is about 10 Gb.

All data on database are in ascii format.

TURKEY

TURKEY

NWP COMPUTER RESOURCES

1. IBM pSeries 690 High Performance Computer: SHT

1 node with 16 CPUs (each 1.3 Ghz)
32 GB total memory size
16x36.4 GB hard disk capacity
AIX Operating System
Workload Manager (WM) is operationaly used.

- MM5 has been run for short-range forecasting with two nests (27km for coarse domain, 9km for inner domain) on SHT since December 2003.
- 23 Vertical levels are in use, it will be upgraded to 36 levels.
- Boundary and initial conditions are provided from ECMWF BC-Suite Project.
- Run Time: 4 times a day operationaly.

Time	Start	Finish
00 UTC	06:20 UTC	08:00 UTC
06 UTC	12:20 UTC	14:00 UTC
12 UTC	18:20 UTC	20:00 UTC
18 UTC	00:20 UTC	02:00 UTC

Forecast Period: t+48, Interval : 3 hour

2. IBM pSeries P630 (Data and Product Server): MEVSIM

4 CPUs (each 1.45 Ghz)
4 GB total memory size
25x36.4 GB hard disk capacity
AIX Operating System

- MEVSIM is our RMDCN primary gateway. It is also used for post processing.
- Metview 3.4 Export Version is run.
- Intranet access for operational use.

Operational use of ECMWF IFS Deterministic Model run outputs.

Horizontal Resolution: 0.5x0.5
Domain: -10.0W- 80.0E/ 60.0N- 30.0N
Forecast Period: 10 days
Parameters: 1. Geopotansial height
2. MSLP,
3. TP,
4. Wind,
5. Temperature, etc.

Operational use of ECMWF EPS Model run outputs.

H. Resolution: ~ 80 km
Forecast Priod: 1-10 days
Products: Probabilities, group means, etc.

Operational use of ECMWF WAVE Model run outputs (Baltic and Mediterranean).

H. Resolution: ~ 27 km
Forecast Period: 5 days
Products: Significant wave height
Mean wave direction
Mean wave period
Swell wave height
Swell wave mean direction
Swell wave mean period

TURKEY

TURKEY

Local Wave Model is run operationaly.

METU3-WAVE model which is originally developed at Middle East Technical Universty-Turkey together Dr Saleh ABDALLA from ECMWF is used to produce daily wave forecasts.

ECMWF Deterministic model run outputs are used as boundary and initial conditions for METU3. It provides forecasts for Black Sea, Marmara and Mediterrean Sea.

H. Resolution:	**0.25*0.25 (~27km)**
Forecast Period:	**T+72**
Products:	**Significant wave height**
	Mean wave direction
	Mean wave period
Interval:	**6 hour**

3. IBM pSeries P630 (with 3-D capability): YAZ
2 CPUs (each 1.45 Ghz)
2 GB total memory size
11x36.4 GB hard disk capacity
AIX Operating System

- YAZ is served as our RMDCN secondary gateway. This machine is also used for as a back up for MEVSIM.
- Metview 3.4 Export Version is run.
- GRADS, NCAR Graphics and RIP graphical software packages are also available for postprocessing.

4. IBM pSeries P630 (Test Machine): TEMMUZ
2 CPUs (each 1.45 Ghz)
2 GB total memory size
4x36.4 GB hard disk capacity
AIX Operating System
INTERNET (ECACCESS) gateway.

5. Intel P4 based workstations (10) run under SuSE Linux 8.2 and Windows XP under VMWare.
3.0 Ghz CPU
72 GB SCSI hard disk capacity
2 GB RAM

- Metview 3.4 Export version is run on desktops.
- NCAR Graphics and RIP are also available on these machines.

6. SGI ORIGIN 2200 Server, R12000 MIPS: SONBAHAR
(300 Mhz x 2 CPU, 1GB memory, 60 GB HDD)
IRIX Operating System

7. SGI ONYX2 Workstation, R10000 MIPS: ILKBAHAR
(180 Mhz x 2CPU, 256 MB memory, 43 GB HDD)
IRIX Operating System

TURKEY

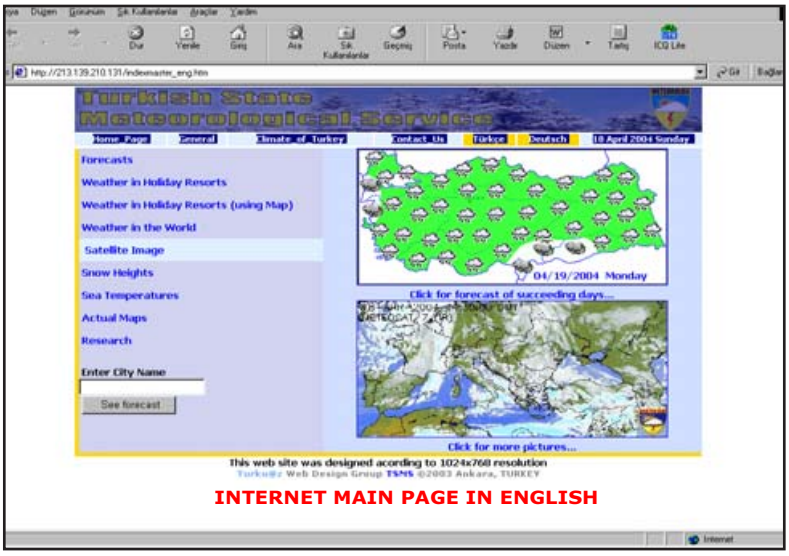
TURKEY

There are 16 ECMWF registered users currently. They all have SecurID tokens.

12 internal (TSMS staff) and 4 external (from universities and governmental institutes)

FUTURE PLANS

- LAN/ATM infrastructure will be upgraded to gigabit infrastructure this year.
- Backup RMDCN (128 Kbps) will be upgraded.
- INTERNET connection will be upgrade to 4Mbps.
- MESSIR software package will be upgrated.
- MESSIR systems will be upgraded.



INTERNET MAIN PAGE IN ENGLISH

UNITED KINGDOM

UNITED KINGDOM

Paul Dando – Met Office

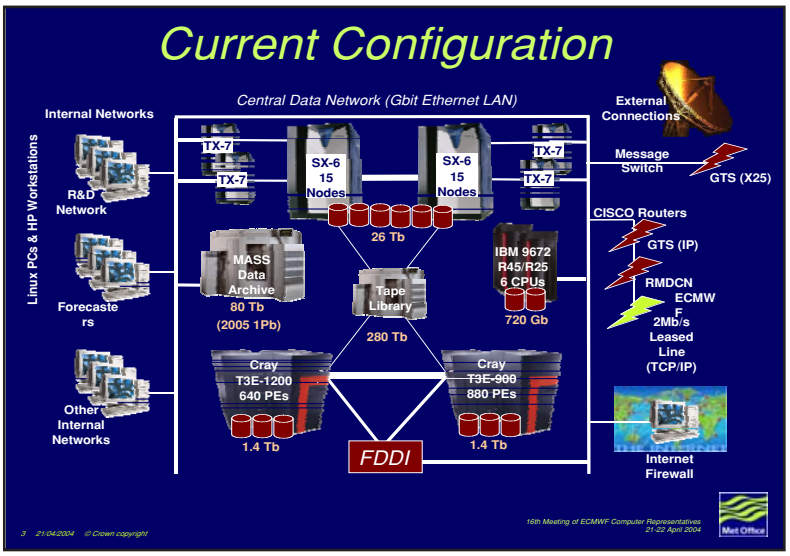
Relocation: Exeter

- Relocation to Exeter completed end 2003
  - 925 people moved from Bracknell (Jun – Dec)
  - Both Cray T3E supercomputers moved
  - Twin 155 Mb/s data link established to facilitate move and maintain business continuity
- First forecast produced in Exeter: 15 Sep 2003
- New building completed: 18 Dec 2003
- Operational service maintained throughout

2 21/04/2004 © Crown copyright

16th Meeting of ECMWF Computer Representatives 21-22 April 2004

Current Configuration



3 21/04/2004 © Crown copyright

16th Meeting of ECMWF Computer Representatives 21-22 April 2004

NEC SX-6 Supercomputer

- First phase accepted end of Jan 2004
  - 30 node NEC SX-6 system split between 2 halls
  - Accessed via 4 NEC TX-7 front ends
  - Provides 6x performance capability of T3Es
  - 26 Tb of disk space forming a GFS across 2 halls
  - Very reliable
- T3Es and SX-6 run in parallel for 3 months
- Operations will switch to SX-6 late Apr 2004
- Second phase scheduled for early 2005
  - Introduction of 15 node SX-6X to double processing power
  - Increase of disk capacity by 50%

4 21/04/2004 © Crown copyright

16th Meeting of ECMWF Computer Representatives 21-22 April 2004

UNITED KINGDOM

UNITED KINGDOM

Desktop Replacement

- Flatscreen PCs used across the Met Office
  - Around 300 Linux desktops for scientists
  - Around 600 Windows XP desktops for other staff
- HP workstations used as servers for compute intensive work
- Increase in network capacity
  - Gigabit Ethernet backbone
  - 100Mb/s to individual desktop

5 21/04/2004 © Crown copyright

16th Meeting of ECMWF Computer Representatives 21-22 April 2004

ECMWF Users

- Users Registered
  - Currently 123 registered users (129 last year)
  - 85 Met Office and 38 UK Universities
- Many users make simple MARS data retrievals
  - Find system easy to use
  - Good documentation
- Few users with large / complex data sets
  - Increased access to ERA-15 & ERA-40 data
    - » Increased load on leased line

6 21/04/2004 © Crown copyright

16th Meeting of ECMWF Computer Representatives 21-22 April 2004

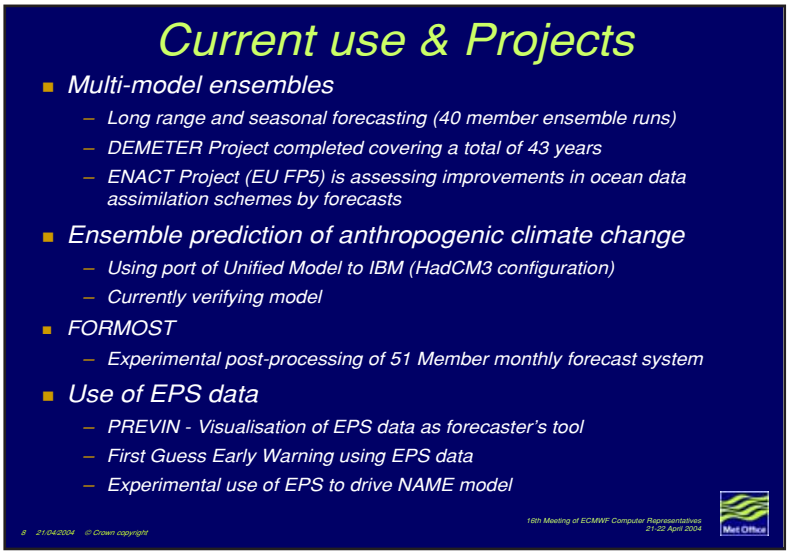
Use of ECMWF Systems

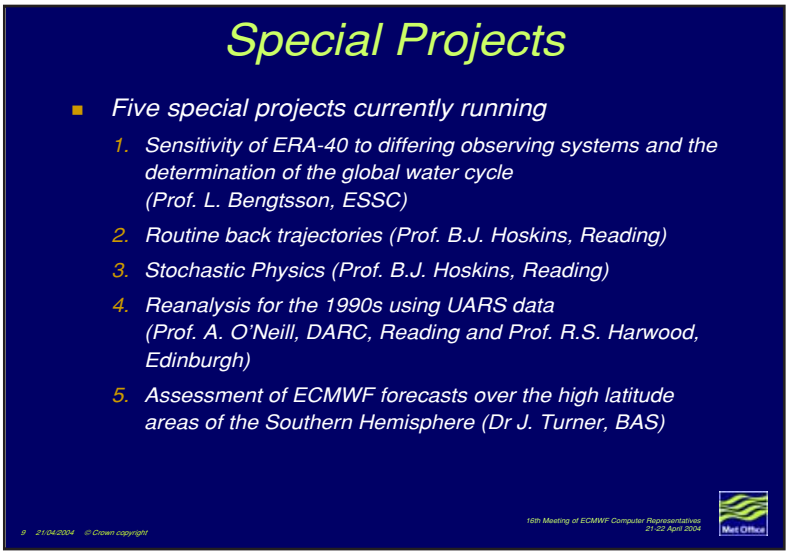
- ECaccess
  - Currently running 4 gateway servers (2 research, 2 operational)
  - All gateways give access via the leased line
  - Some problems – improved stability in recent months
- Transition to new ECgate service underway
- Metview
  - Greatly improved response with the new Linux desktop
  - Many local macros
  - Automated MARS retrievals
- Use of HPCF
  - Unified Model ported to IBM (using MPP code)
  - Used 32% of total SBU allocation in 2003

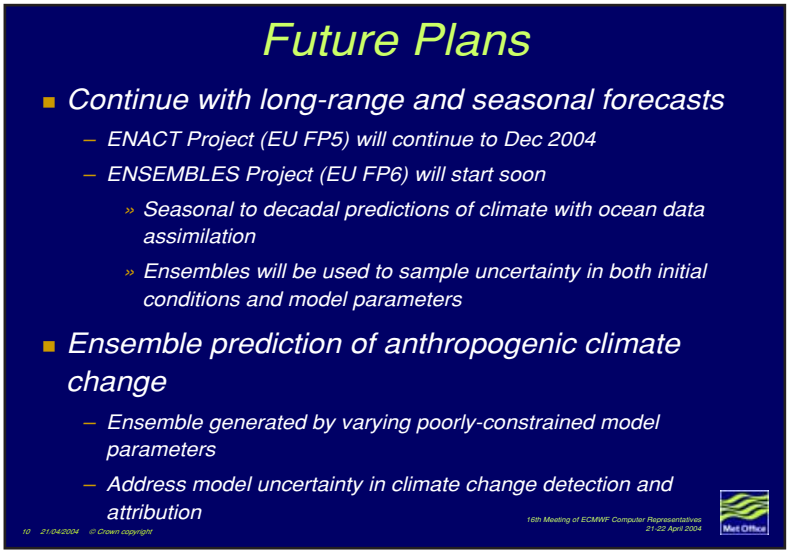
7 21/04/2004 © Crown copyright

16th Meeting of ECMWF Computer Representatives 21-22 April 2004

UNITED KINGDOM

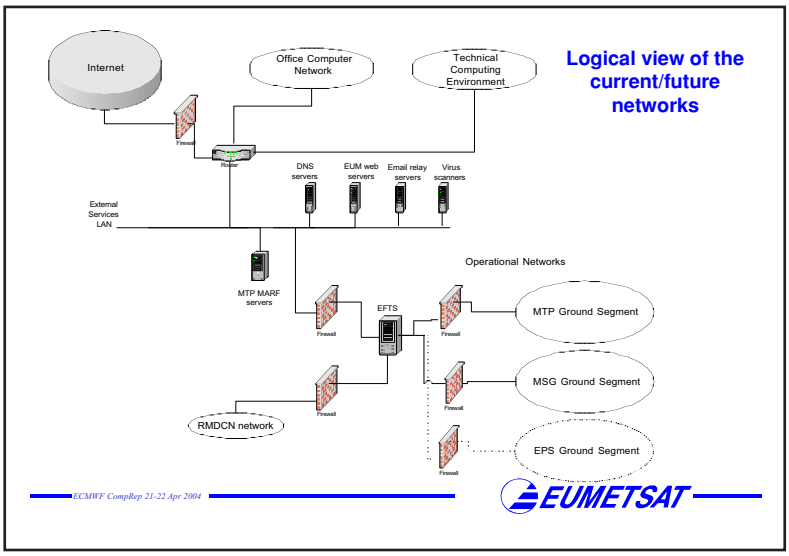






EUMETSAT **EUMETSAT**

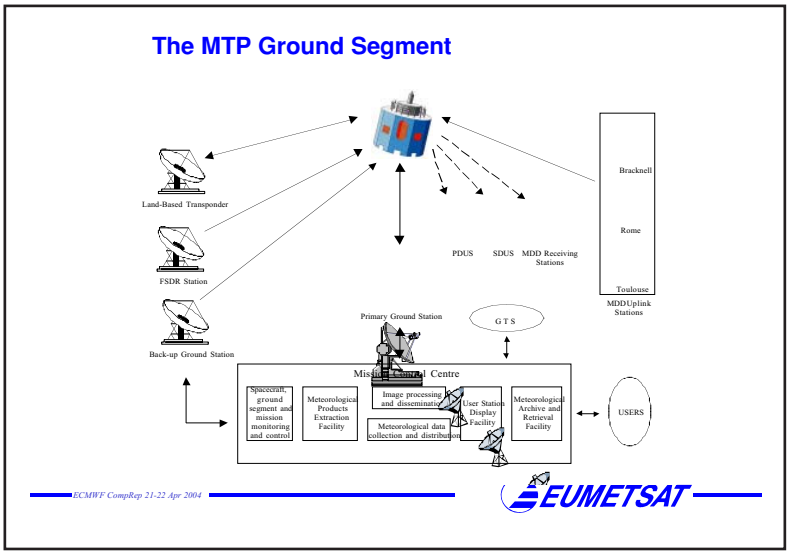
Martin Dillmann – EUMETSAT



Meteosat Transition Program

- Relocation to new control centre finished in 2003
- Installation of an extended backup control centre in Fucino with full mission capabilities is still ongoing - to be finished by mid 2004
- Replacement of the terrestrial link (768 MBit/s) to Fucino with an E1 (2 MBit/s) link plus primary ISDN backup line

ECMWF CompRep 21-22 Apr 2004 EUMETSAT



Meteosat Second Generation (1/2)

- Operations started by end of 2003
- Some facilities still in W&M phase
- Re-engineering and upgrade cycle of Hardware and Software has begun
- New Gigabit-Ethernet network infrastructure is currently being deployed

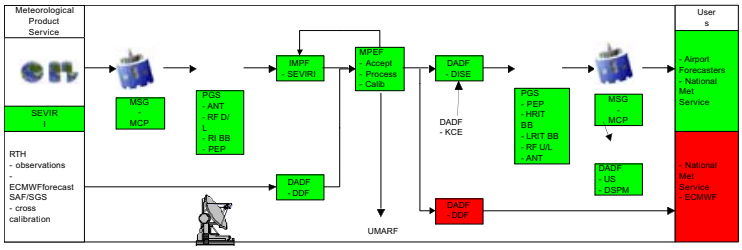
ECMWF CompRep 21-22 Apr 2004 EUMETSAT

Meteosat Second Generation (2/2)

- Finalisation of DVB based dissemination facilities - EUMETCAST
- Installation of SAN infrastructure serving all ground segment facilities
- Re-design of the archive facility (UMARF) using SAN infrastructure and LTO-2 tape silos

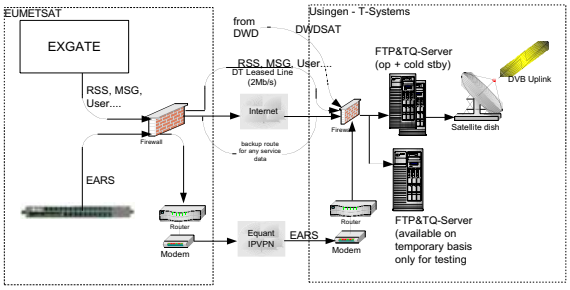
ECMWF CompRep 21-22 Apr 2004 EUMETSAT

MSG Meteorological Products Service Chain Flow Diagram (original design)



ECMWF CompRep 21-22 Apr 2004 EUMETSAT

EUMETCAST implementation



ECMWF CompRep 21-22 Apr 2004 EUMETSAT

Technical computing environment (1/3)

- New non-operational system for cpu- and data-intensive computations:
  - MSG MPEF algorithm development
  - EPS simulation
  - EPS algorithm prototyping
- Multi-platform system with 100 TB storage capacity

ECMWF CompRep 21-22 Apr 2004 EUMETSAT

Technical computing environment (2/3)

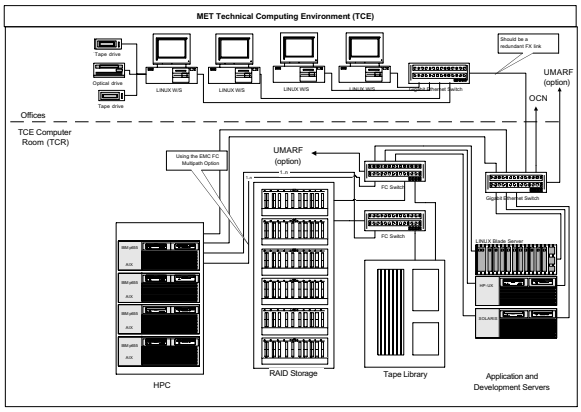
- Computing:
  - IBM 8-way p655 based HPC system
  - HP 5470, SUN V880 application server
- Storage:
  - EMC Clariion CX600 RAID Array (~ 50 TB max)
  - OVERLAND Neo 4xxx tape library (max. 48 TB native)
  - GPFS file system plus TSM for backup and disk extension
- SAN:
  - Switched Fibre Channel (2 Gb/s), redundant
  - Veritas SAN Point Control as management tool

ECMWF CompRep 21-22 Apr 2004 EUMETSAT

Technical computing environment (3/3)

- Network:
  - Private network
  - Switched Gigabit Ethernet Network (SX and 1000-T)
  - HP Procurve switches
- User Workstation:
  - Standard PCs with Linux as user front-ends
  - Gigabit Ethernet connection

ECMWF CompRep 21-22 Apr 2004 EUMETSAT



ECMWF CompRep 21-22 Apr 2004 EUMETSAT

F. Hofstadler asked when EUMETCAST would be fully operational and if there were plans for MDD dissemination.

M. Dillmann replied that EUMETCAST was operational, though he was not sure whether the encryptional path was operational. Some testing of MDD dissemination was underway.

ANNEX 1 ANNEX 1

Sixteenth Meeting of Computing Representatives

ECMWF, Shinfield Park, Reading, U.K., 21-22 April 2004

Participants

Austria	Gerhard Hermann
Belgium	Liliane Frappez
Czech Republic	Karel Ostatnicky Karel Pesata
Denmark	Thomas Lorenzen
Finland	Kari Niemelä
France	Marion Pithon
Germany	Elisabeth Krenzien
Greece	Ioannis Mallas
Hungary	Laszlo Tölgyesi
Ireland	Paul Halton
Italy	Giuseppe Tarantino
Netherlands	Hans De Vries
Norway	Rebecca Rudsar
Romania	Elena Toma
Serbia & Montenegro	Vladimir Dimitrijevic
Slovenia	Miha Razinger
Spain	Eduardo Monreal
Sweden	Rafael Urrutia
Switzerland	Peter Roth
Turkey	Bülent Yagci
United Kingdom	Paul Dando
Eumetsat	Martin Dillmann
ECMWF:	Sylvia Baylis
	Petra Berendsen
	Jens Daabeck
	Matteo Dell'Acqua
	Richard Fisker
	Helene Garçon
	Laurent Gougeon
	John Greenaway
	Alfred Hofstadler
	Norbert Kreitz
	Dominique Lucas
	Carsten Maass
	Umberto Modigliani
	Pam Prior
	Baudouin Raoult
	Deborah Salmond
	Neil Storer
	Walter Zwieflhofer

ANNEX 2 ANNEX 2

Programme

Wednesday, 21 April

09.30 Coffee
10.00 Welcome
ECMWF's computer status and plans ..... W. Zwieflhofer
11.00 Member States and Co-operating States presentations
12.30 Lunch
14.00 Member States and Co-operating States presentations (continued)
14.30 HPCF and DHS update ..... N. Storer
14:50 Early experience on Phase3 test system ..... D. Salmond
15.10 Data and Services update ..... B. Raoult
15.30 Graphics update ..... J. Daabeck
15.45 ECaccess status ..... L. Gougeon
16.00 Coffee
16:30 Linux cluster presentations and discussion ..... MS/Co-op + ECMWF
17.30 Cocktails
18:30 Transport to Hotels
19:45 Informal dinner at Pepe Sale Restaurant

Thursday, 22 April

09.00 Member States and Co-operating States presentations (continued)
10.30 Coffee
11:00 User registration: update and demo ..... P. Kogel
11:20 Web access control changes ..... C. Valiente
11.40 Survey of external users and status of ecgate migration ..... U. Modigliani
12.00 Discussion
12.30 End of meeting
13:00 Transport to Heathrow