The Norwegian Metacenter for Computational Science - Notur II

Jacko Koster

www.notur.no
1986 - 1991: First HPC project (NTH, SINTEF)
1988 - : Met. office starts using the HPC resources

1991 - 1995: Second HPC project (NTH, UiO, UiB)
Nordic cooperation

1995 - 2000: Third HPC project (NTNU, UiO, UiB)
Central resource allocations

2000 - 2004: Fourth HPC project (Notur I)
Partners: NTNU, UiO, UiB, UiT, met.no, Statoil, SINTEF, Ceetron; Budget 22 MNOK RCN

2005 – 2014: Fifth HPC project (Notur II)
Partners: UNINETT Sigma, NTNU, UiO, UiB, UiT, met.no. Budget 22 MNOK RCN

From January 2006, Notur is a project under eVITA

2006 – 2015 eVITA programme
Notur is the national infrastructure project for computational science in Norway

The project provides resources and services to
- education and research at universities and colleges
- operational weather forecasting and research at the Meteorological institute
- research and engineering at research institutes and industry who contribute to the project

Project duration: 2005-2014
Budget 2005: 21.7 MNOK from RCN
2006: 21.7 + 3.0 MNOK from RCN
- Provide a powerful and cost-effective infrastructure for computational science
- Enable efficient utilization of the infrastructure
- Develop a national grid for computation and data
- Establish international collaboration in infrastructure and computational science
- Disseminate computational science as an important discipline
infrastructure = hardware + software + services + support

Computers
From desktop to HPC …

Tools
basic utilities, authorization, authentication, …

End-user functionality
interfaces, APIs, …

Operations
Monitoring, tuning, maintenance, upgrades, …

Storage
Disk, tape, …

Grid
connect hardware, middlewares, …

Performance
responses, turnaround, throughput, efficiency, optimal resource utilization

Support
Basic (helpdesk), specialized/advanced, …

Networks
Mbps, Gbps, lambda, …

Quality Assurance
Access, reliability, predictability, availability, fault-tolerance, …

Instruments
Scientific, medical, sensor, …

User & Application
INFRASTRUCTURE

infrastructure = hardware + software + services + support

Computers
From desktop to HPC …

Storage
Disk, tape, …

Networks
Mbps, Gbps, lambda, …

Instruments
Scientific, medical, sensor, …

Tools
basic utilities, authorization, authentication, …

End-user functionality
interfaces, APIs, …

Performance
responses, turnaround, throughput, efficiency, optimal resource utilization

Quality Assurance
Access, reliability, predictability, availability, fault-tolerance, …

Grid
connect hardware, middlewares, …

Operations
Monitoring, tuning, maintenance, upgrades, …

Support
Basic (helpdesk), specialized/advanced, …

Notur = infrastructure for computational science

User & Application
A **national** infrastructure for computational science to ...

- coordinate investments in hardware and software
- provide heterogeneous resources
- allow aggregation/sharing of resources
- allow aggregation/sharing of competence
- provide cost-efficient solutions
- provide nation-wide sustained services (standardization, interoperation, harmonization, policies)
- ...

INFRASTRUCTURE
Why HPC?

Why computational science?
- Replace field/laboratory experiments by simulation
- Explore territory where experiments are impossible
- Validate theory
- Predict

Why high-performance computers?
- For time-critical applications, e.g., weatherforecast
- Solve large-scale compute-/data-intensive problems
- Moore’s Law: performance doubles every 18 months at constant cost. In 10 years ca. 100x increase.
- I.e., in 10 years, a PC has caught up with a 100 CPU machine
- Or, a scientist that can efficiently use a 100 CPU machine is 10 years ahead of a scientist using working on a single PC
<table>
<thead>
<tr>
<th>Year</th>
<th>Type</th>
<th>#CPUs/cores</th>
<th>Peak</th>
<th>Memory</th>
<th>Disk</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001</td>
<td>IBM p690</td>
<td>96</td>
<td>499 Gflop/s</td>
<td>320 GB</td>
<td>6 TB</td>
<td>22 Mkr</td>
</tr>
<tr>
<td>2006</td>
<td>IBM p575</td>
<td>992</td>
<td>7 Tflops/s</td>
<td>2 TB</td>
<td>70 TB</td>
<td>30 Mkr</td>
</tr>
<tr>
<td></td>
<td>Off-the-shelf-technology:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>cluster node</td>
<td>4</td>
<td>20 Gflops/s</td>
<td>8 GB</td>
<td>250 GB</td>
<td>0.02 Mkr</td>
</tr>
<tr>
<td>2007</td>
<td>cluster</td>
<td>4000</td>
<td>20 Tflops/s</td>
<td>8 TB</td>
<td>250 TB</td>
<td>20 Mkr</td>
</tr>
</tbody>
</table>
Application profile:

- Geophysics, fluid dynamics
- Nanotechnology, chemistry
- Bio-informatics, high energy physics
### Usage 2005-2006

**Oct 1, 2005 – Sept. 30, 2006**

<table>
<thead>
<tr>
<th>Organisation</th>
<th>Discipline</th>
<th>Funding</th>
</tr>
</thead>
<tbody>
<tr>
<td>met.no</td>
<td>geophysics</td>
<td>NORKLIMA 16%</td>
</tr>
<tr>
<td>UiO</td>
<td>chemistry</td>
<td>University 16%</td>
</tr>
<tr>
<td>UiT</td>
<td>physics</td>
<td>SFF / SFI 10%</td>
</tr>
<tr>
<td>UiB</td>
<td></td>
<td>EU 9%</td>
</tr>
<tr>
<td>NERSC</td>
<td></td>
<td>NANOMAT 8%</td>
</tr>
<tr>
<td>NTNU</td>
<td></td>
<td>FRINAT 7%</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td>YFF 6%</td>
</tr>
</tbody>
</table>

Between 55-65 active projects

Ca. 200 TB archived data (mostly climate)
Application profile:

- geophysics, fluid dynamics
- nanotechnology, chemistry
- bio-informatics, high energy physics
Application profile:

non-scalable

embarrassingly parallel

- geophysics, fluid dynamics
- nanotechnology, chemistry
- bio-informatics, high energy physics

non-scalable : bandwidth / latency bound applications
embarrassingly parallel : many independent tasks
Hardware profile:

- geophysics, fluid dynamics
- nanotechnology, chemistry
- bio-informatics, high energy physics

Job volume

Process coupling

- embarrassingly parallel
- non-scalable

☐: one solution that fits all – not cost-efficient
Hardware profile:

- geophysics, fluid dynamics
- nanotechnology, chemistry
- bio-informatics, high energy physics

Job volume

Process coupling

☐ ☐ ☐ : cover what is needed (multiple machines)
Hardware profile:

- **geophysics, fluid dynamics**
- **nanotechnology, chemistry**
- **bio-informatics, high energy physics**

**Job volume**

**Capacity**:
- Blue: clusters with default node interconnect (embarrassingly parallel, MPI)
- Green: clusters with fast node interconnect (MPI)
- Red: capability: massively parallel platforms (MPI), shared-memory (OpenMP)
Performance pyramid:

- **Capacity**:
  - Blue: Clusters with default node interconnect (embarrassingly parallel, MPI)
  - Green: Clusters with fast node interconnect (MPI)
  - Red: Capability: massively parallel platforms (MPI), shared-memory (OpenMP)

- **Process Coupling**
  - Njord
  - Snowstorm

- **Job Volume**
  - Expensive (5000 Euro/CPU)
  - Cheap (600 Euro/CPU)
NTNU

njord.hpc.ntnu.no

992 power5+ CPUs, 2 TB memory, 70 TB disk

7 Tflop/s

11/2006
The Snowstorm Cluster

HP rx4640: 408 Itanium2 CPUs, 408 GB Memory, 30 TB Disk
Infiniband interconnect, ROCKS Cluster Distribution

2.3 Tflop/s
Status:

**snowstorm**

- Run: 382
- Susp: 0
- Pend: 255
- Block: 96
- Total CPUs: 418

**njord**

- Run: 286
- Susp: 0
- Pend: 150
- Block: 0
- Total CPUs: 896

Taken from www.notur.no
Tue Jan 30 2007 9.30 am
<table>
<thead>
<tr>
<th>Partner</th>
<th>System</th>
<th>CPU</th>
<th>#CPUs</th>
<th>#Gflops</th>
<th>until</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTNU</td>
<td>SGI Origin 3800</td>
<td>r14000</td>
<td>896</td>
<td>1000</td>
<td>12/2006</td>
</tr>
<tr>
<td>UIB</td>
<td>IBM p690 Regatta</td>
<td>power4</td>
<td>96</td>
<td>499</td>
<td>03/2007</td>
</tr>
<tr>
<td>UIB</td>
<td>IBM 1300 cluster</td>
<td>Pentium</td>
<td>64</td>
<td>80</td>
<td>03/2007</td>
</tr>
<tr>
<td>UIO</td>
<td>HP SuperDome</td>
<td>Itanium2</td>
<td>64</td>
<td>384</td>
<td>09/2007</td>
</tr>
<tr>
<td>UIT</td>
<td>HP rx4640</td>
<td>Itanium2</td>
<td>418</td>
<td>2300</td>
<td>09/2008</td>
</tr>
<tr>
<td>NTNU</td>
<td>IBM p575+</td>
<td>power5+</td>
<td>992</td>
<td>7000</td>
<td>11/2010</td>
</tr>
</tbody>
</table>

Expansion 01/2007 for WLCG, ca 120 CPUs, 135 TB storage
Expansion 2H2007 with new resource

<table>
<thead>
<tr>
<th>Local clusters</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>UiB</td>
<td>IBM e1350 cluster</td>
<td>AMD</td>
<td>172</td>
<td>825</td>
</tr>
<tr>
<td>UiO</td>
<td>Dell cluster</td>
<td>Xeon</td>
<td>384</td>
<td>2200</td>
</tr>
<tr>
<td>NTNU</td>
<td>Dell cluster</td>
<td>Xeon</td>
<td>64</td>
<td></td>
</tr>
</tbody>
</table>
New investments:
- User needs
- What type of hardware is already in operation
- Vendor roadmaps
- Available funding
- Location of the resource, total size of machine park, …

Trends in hardware:
- Heat problems prevent increase in clock-speed: multi-core
- Hardware investments in coming years will have > 1000 cores

Software:
- Problem/Model sizes increase
- Need to improve scalability of software
Application support: by investing x Mkr in competence, one can save y Mkr on hardware, x<y. More realistic, one can use existing resources for more/larger computations, larger/finer models.
Aim is to provide high-level assistance
- enable scientist to focus (better) on science
- increase job throughput / scientific productivity
- optimize utilization of expensive resources
- complex application enabling

Targets
- User groups with large computational/storage demands
- Groups with no experience in using high-end resources

Examples: scalability, performance improvement, usability, software development (limited)

September 2006: 22 applications received, 11 approved
February 22, 2007: deadline on-going call
Bjørn Angelsen, Department of Circulation and Imaging, NTNU
- 3D nonlinear wave propagation in heterogeneous media in ultrasound

Laurent Bertino, Nansen Environmental and Remote Sensing Center
- Reducing the memory requirements of EnKF and HYCOM-NORWECOM

Tore Børvik, Department of Structural Engineering, NTNU
- Tuning LS-DYNA on the new IBM system in Trondheim

Finn Drabløs, Department of Cancer Research and Molecular Medicine, NTNU
- Optimisation and porting of code for OrthoMCL

Helge Drange, Bjerknes Centre for Climate Research, UiB
- Adaptation of the Bergen Climate Model to cluster type computer systems

Olav Holberg, Holberg Research AS
- Efficient utilization of HPC clusters for the ANIVEL project

Trond Iversen, Department of Geosciences, UiO
- Porting the CAM-Oslo model to new computers

Signe Kjelstrup, Department of Chemistry, NTNU
- Parallelization of the forces calculation in ZEO-GAS

Arvid Lundervold, Department of Biomedicine, UiB
- Quantitative brain imaging in health and disease

Knut Helge Midtbø, Meteorological Institute
- Improved quality and unparalleled regularity of met.no operational atmospheric forecasts

Bjørnar Pettersen, Department of Marine Technology, NTNU
- Improving the parallel performance of the CFD code Vista
Performance pyramid:

- **Capacity**: clusters with default node interconnect (embarrassingly parallel, MPI)
- **Capacity**: clusters with fast node interconnect (MPI)
- **Capability**: massively parallel platforms (MPI), shared-memory (OpenMP)

---

eVITA forskermøte 30-01-2007
European performance pyramid:

- **Tier 0**: few very large capability systems in Europe
- **Tier 1**: national HPC centres
- **Tier 2**: local machines

**Job volume**

- Machine

**Process coupling**

- Blue: capacity: clusters with default node interconnect (embarrassingly parallel, MPI)
- Green: capacity: clusters with fast node interconnect (MPI)
- Red: capability: massively parallel platforms (MPI), shared-memory (OpenMP)
DATA

Data stored ’right next’ to facility where data is needed/generated

→ Explicit duplication of data by user (data management by user)
→ Machine-specific data formats (portability issues)

Ca. 300 TB data archived (mostly climate)

Better attention to storage:
1. Decoupling of data/storage from computer → distributed solutions
2. Optimize location of data (repositories, replication, …)
3. Long-term storage of data (surviving shifts in technology)

Better attention to storage management coincides with upgrade of the national research network from 2.5 Gbit/s to 10 Gbit/s
Grid is concerned with connecting resources and providing end-user functionalities for **resource scheduling** ('moving applications around') and **data management** ('moving data around') for a variety of disciplines.

**NorGrid:**
- Deploy and operate a reliable grid on existing infrastructure
- Provide bigger and better resources than provided by a single machine
- Deploy services for sharing, collecting, retrieving data and distributed data management

Close coupling with Notur and Nordic, European projects: NDGF, SweGrid, M-grid, KnowARC, EGEE, SIRENE, FEIDE, GigaCampus, …
Characterization (Foster):
- Coordinates resources that are not subject to centralized control
- Uses standard open general-purpose protocols and interfaces
- Delivers non-trivial qualities of service.

Keywords: sustainability, availability, security, scalability, performance, (meta)data management, resource discovery, information services

Virtualization: decouple hardware resources from the applications running on it: allows dynamic and flexible management of work load and data

1. Job brokering: user submits job to ‘the grid’ (with well-defined constraints)

2. Location of data transparent to application: physical location of data is replaced by logical location: spread out data is visible as a regular homed directory; fast staging techniques to get on time data to/from the application

3. Advanced services to collect, gather, share and publish data
Job brokering for resource scheduling:

NorGrid – Norwegian Grid Infrastructure

雪暴雪（snowstorm）

 njord

Run: 286  Susp: 0  Pend: 150
Block: 0  Total CPUs: 896

Taken from www.notur.no
Tue Jan 30 2007 9.30 am
EGI - European Grid Infrastructure

EU infrastructure / EGI

Region 1 / RGI

NorGrid

Region N / NDGF

NGI A

Univ. A1
Org. A2
Inst. A3

NGI B

Univ. B1
Org. B2
Inst. B3

NGI X

Univ. X1
Org. X2
Inst. X3

NDGF

EGI - European Grid Infrastructure

User A

EU infrastructure / EGI

Region 1 / RGI

NGI A

Univ. A1

Org. A2

Inst. A3

NGI B

Univ. B1

Org. B2

Inst. B3

NorGrid

Univ. C1

Org. C2

Inst. C3

User C

Region N / NDGF

NGI X

Univ. X1

Org. X2

Inst. X3

User D

User B
The Large Hadron Collider (LHC) at CERN is a particle accelerator which will probe deeper into matter than ever before. Due to switch on in 2007, it will ultimately collide beams of protons at an energy of 14 TeV. Beams of lead nuclei will be also accelerated, smashing together with a collision energy of 1150 TeV.

1 TeV is roughly the energy of motion of a flying mosquito... What makes the LHC so extraordinary is that it squeezes energy into a space about a million million times smaller than a mosquito…

LHC uses superconductivity. LHC operates at 300 degrees below room temperature. Data will be collected as of April 2008.
LHC accelerator

Five experiments: ATLAS, ALICE, CMS, LHCb, TOTEM
WLCG - Hierarchy

- Tier-0 (CERN)
  - Records RAW data (1.25 GB/s ALICE)
  - Distributes second copy to Tier-1s
  - Calibrates and does first-pass reconstruction

- Tier-1 centres (11 defined)
  - Manages permanent storage – RAW, simulated, processed
  - Capacity for reprocessing and analysis

- Tier-2 centres (~ 100 identified)
  - Monte Carlo event simulation
  - End-user analysis

- Tier-3 (hundreds, local resources)
Any Tier-2 may access data at any Tier-1

Tier-2s and Tier-1s are inter-connected by the general purpose research networks
WLCG - CPU requirements

Year

2007 2008 2009 2010

MSI2000

LHCb-Tier-2
CMS-Tier-2
ATLAS-Tier-2
ALICE-Tier-2
LHCb-Tier-1
CMS-Tier-1
ATLAS-Tier-1
ALICE-Tier-1
LHCb-CERN
CMS-CERN
ATLAS-CERN
ALICE-CERN
WLCG - Disk requirements

![Graph showing disk requirements for different experiments from 2007 to 2010. The graph compares LHCb-Tier-2, CMS-Tier-2, ATLAS-Tier-2, ALICE-Tier-2, LHCb-Tier-1, CMS-Tier-1, ATLAS-Tier-1, ALICE-Tier-1, LHCb-CERN, CMS-CERN, ATLAS-CERN, and ALICE-CERN. The y-axis represents PB (Petabytes), and the x-axis represents the year from 2007 to 2010. The graph shows a steady increase in disk requirements over the years.]
WLCG - Tape requirements

Year

PB

2007 2008 2009 2010

LHCb-Tier-1
CMS-Tier-1
ATLAS-Tier-1
ALICE-Tier-1
LHCb-CERN
CMS-CERN
ATLAS-CERN
ALICE-CERN
LHC timeline

- Sep05 - SC3 Service Phase
- June06 – SC4 Service Phase
- Sep06 – Initial LHC Service in stable operation
- Apr07 – LHC Service commissioned

2005 2006 2007 2008

- SC3
- SC4

LHC Service Operation

- cosmics
- First beams
- First physics
- Full physics run
The Nordic Data Grid Facility is a collaboration between Denmark, Finland, Norway, Sweden.

Motivation: to ensure that researchers in the Nordic countries can create and participate in computational challenges of scope and size unreachable for the national research groups alone.

NDGF is a production grid facility that leverages existing, national computational resources and grid infrastructures.

The first ‘customer’ is the High Energy Physics community through the operation of the Nordic Tier-1.

Legal entity of NDGF is NORDUnet (Kastrup) … www.ndgf.org
Nordic Tier-1

Phase I 2003-2005, pilot project with four postdocs
Phase II started June 2006, prepare for LHC experiments

Partners in Tier-1:
- CSC, Espoo, Finland
- Niels Bohr Institute, Copenhagen, Denmark
- NSC, Linköping, Sweden
- HPC2N, Umeå, Sweden
- PDC, KTH, Stockholm, Sweden
- University of Bergen, Norway
- University of Oslo, Norway

Additional Tier-2 partners to be defined
Nordic Tier-1 commitments for LHC (ALICE/ATLAS/CMS)

<table>
<thead>
<tr>
<th></th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU (kSI2K)</td>
<td>520</td>
<td>1340</td>
<td>2610</td>
<td>3470</td>
<td>4280</td>
</tr>
<tr>
<td>Disk (Tbytes)</td>
<td>160</td>
<td>440</td>
<td>890</td>
<td>1500</td>
<td>2200</td>
</tr>
<tr>
<td>Tape (Tbytes)</td>
<td>240</td>
<td>435</td>
<td>872</td>
<td>1500</td>
<td>2260</td>
</tr>
<tr>
<td>WAN (Mbits/sec)</td>
<td>2000</td>
<td>5000</td>
<td>10000</td>
<td>20000</td>
<td>20000</td>
</tr>
</tbody>
</table>
Norwegian contribution is derived from 20% of Nordic Tier-1:

<table>
<thead>
<tr>
<th></th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU (kSI2K)</td>
<td>264</td>
<td>528</td>
<td>694</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disk (Tbytes)</td>
<td>75</td>
<td>200</td>
<td>325</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tape (Tbytes)</td>
<td>60</td>
<td>160</td>
<td>260</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WAN (Mbits/sec)</td>
<td>2500</td>
<td>10000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

UiB and UiO host Tier-1 resources for Norway

Definition of Norwegian Tier-2 in April 2007
Further information

www.notur.no
sigma@uninett.no