The Norwegian Metacenter for Computational Science - Notur II

Jacko Koster

www.notur.no

HPC projects in Norway

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

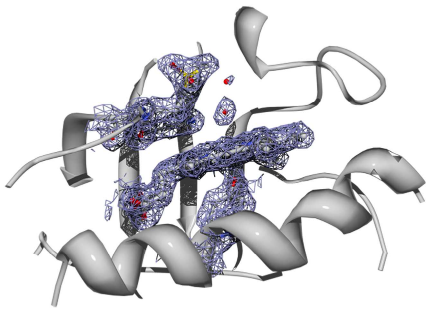
2006 – 2015 eVITA programme

today

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 weatherforecasting and research at the Meteorological institute
- research and engineering at research institutes and industry who contribute to the project



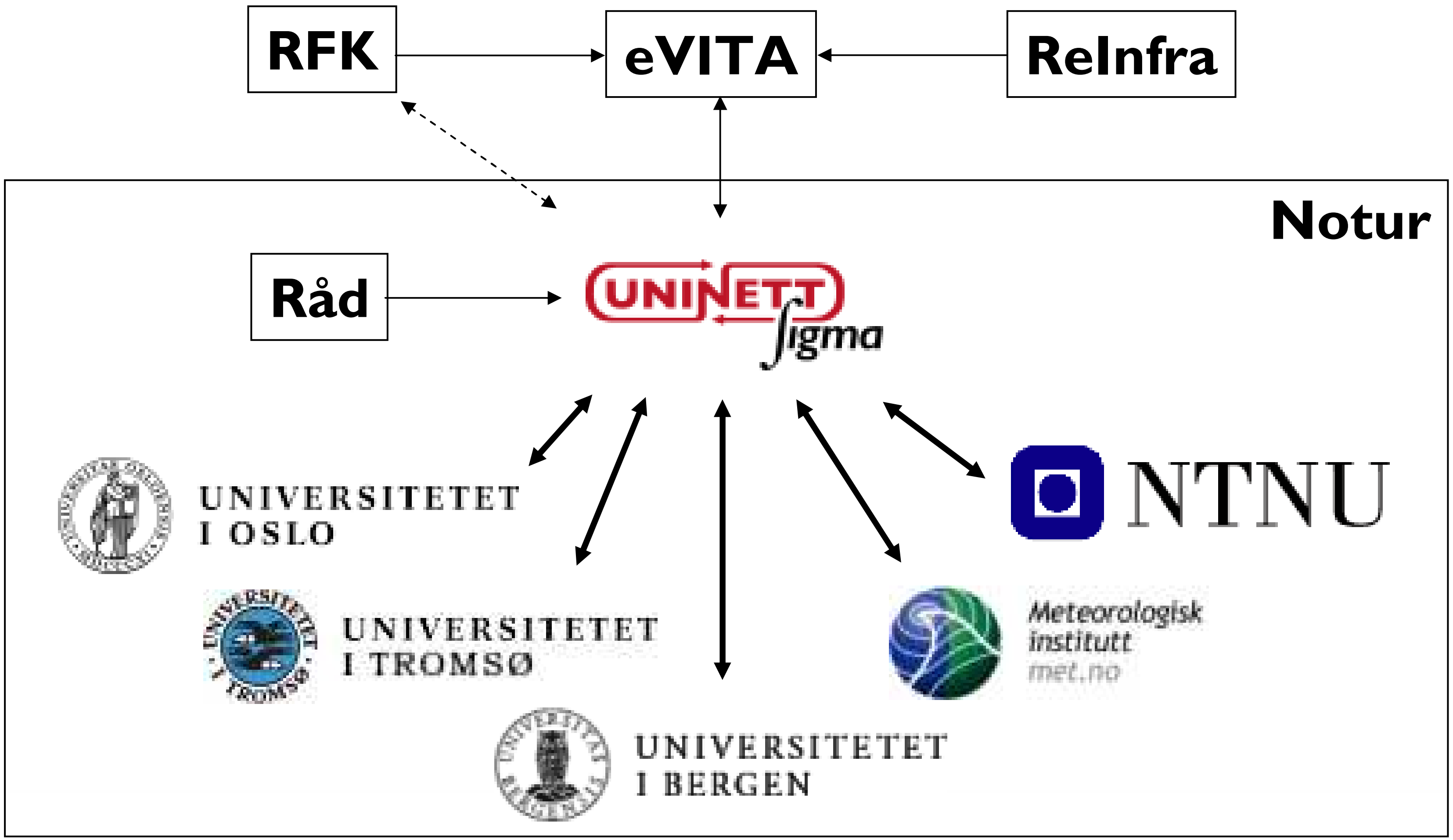
Project duration: 2005-2014

Budget 2005: 21.7 MNOK from RCN

2006: 21.7 + 3.0 MNOK from RCN

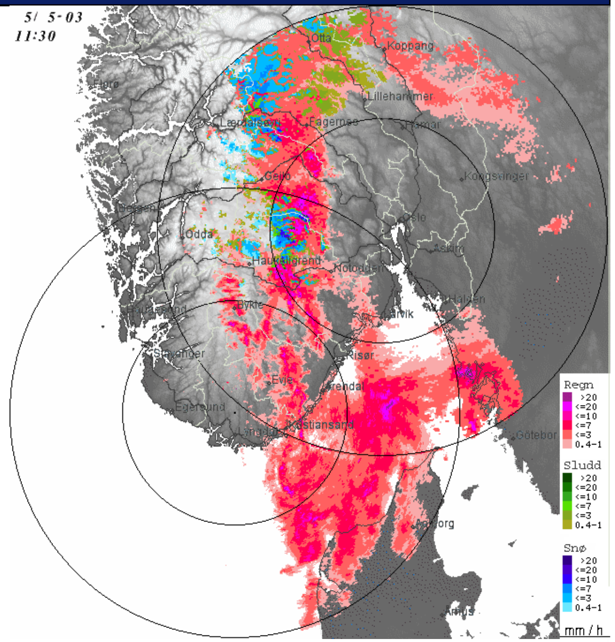
partners

Norges forskningsråd

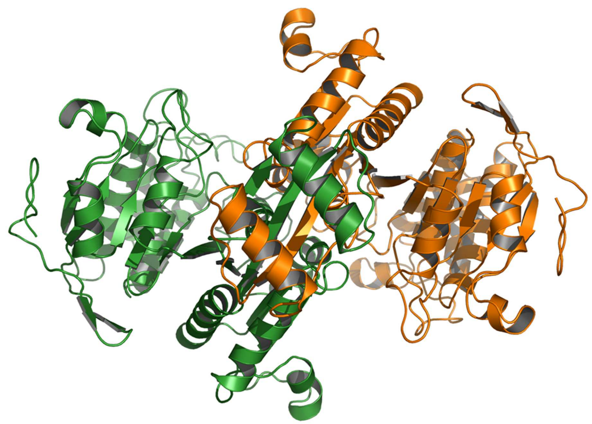


partial goals

- 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

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, …

Grid
connect hardware, middlewares, …

End-user functionality
interfaces, APIs, …

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

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

Operations
Monitoring, tuning, maintenance, upgrades, …

Support
Basic (helpdesk), specialized/advanced, …

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, …

Grid
connect hardware, middlewares, …

End-user functionality
interfaces, APIs, …

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

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

Operations
Monitoring, tuning, maintenance, upgrades, …

Support
Basic (helpdesk), specialized/advanced, …

User & Application

Notur = infrastructure for computational science

INFRASTRUCTURE

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)
- ...

Why HPC?

Why computational science?

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

Why high-performance computers?

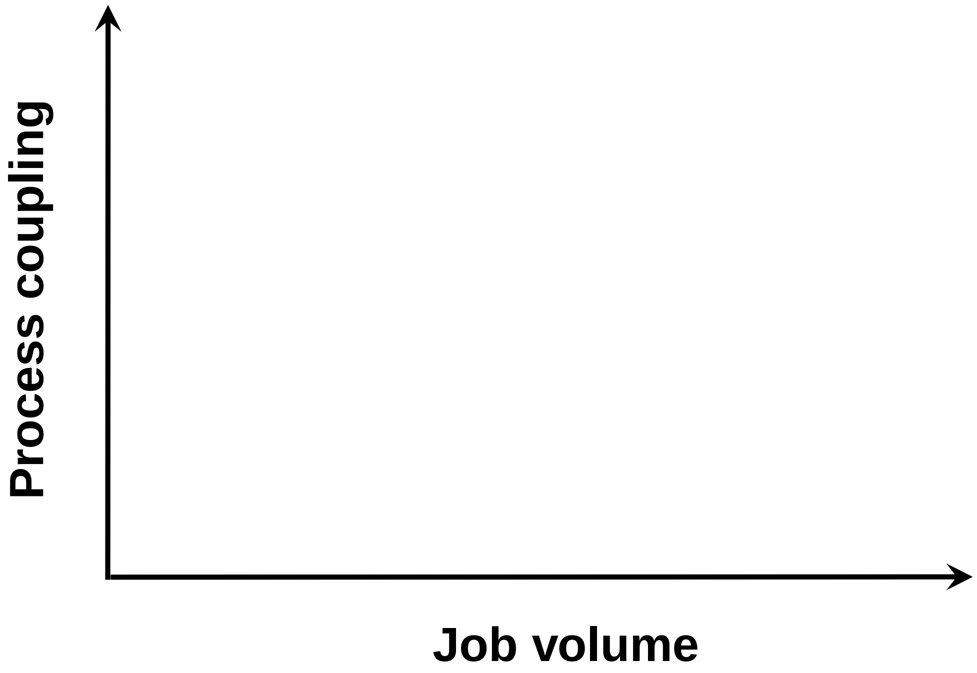
- For time-critical applications, e.g., weatherforecast
- Solve large-scale compute-/data- intensive problems
- Moore's Law: performance doubles every 18 months at constant cost. In 10 years ca. 100x increase.
- I.e., in 10 years, a PC has caught up with a 100 CPU machine
- Or, a scientist that can efficiently use a 100 CPU machine is 10 years ahead of a scientist using working on a single PC

Year	Type	#CPUs/ cores	Peak	Memory	Disk	Cost
1996	Origin 2000	128	51 Gflop/s	24 GB	700 GB	25 Mkr
2001	IBM p690	96	499 Gflop/s	320 GB	6 TB	22 Mkr
2006	IBM p575	992	7 Tflops/s	2 TB	70 TB	30 Mkr

Off-the-shelf-technology:

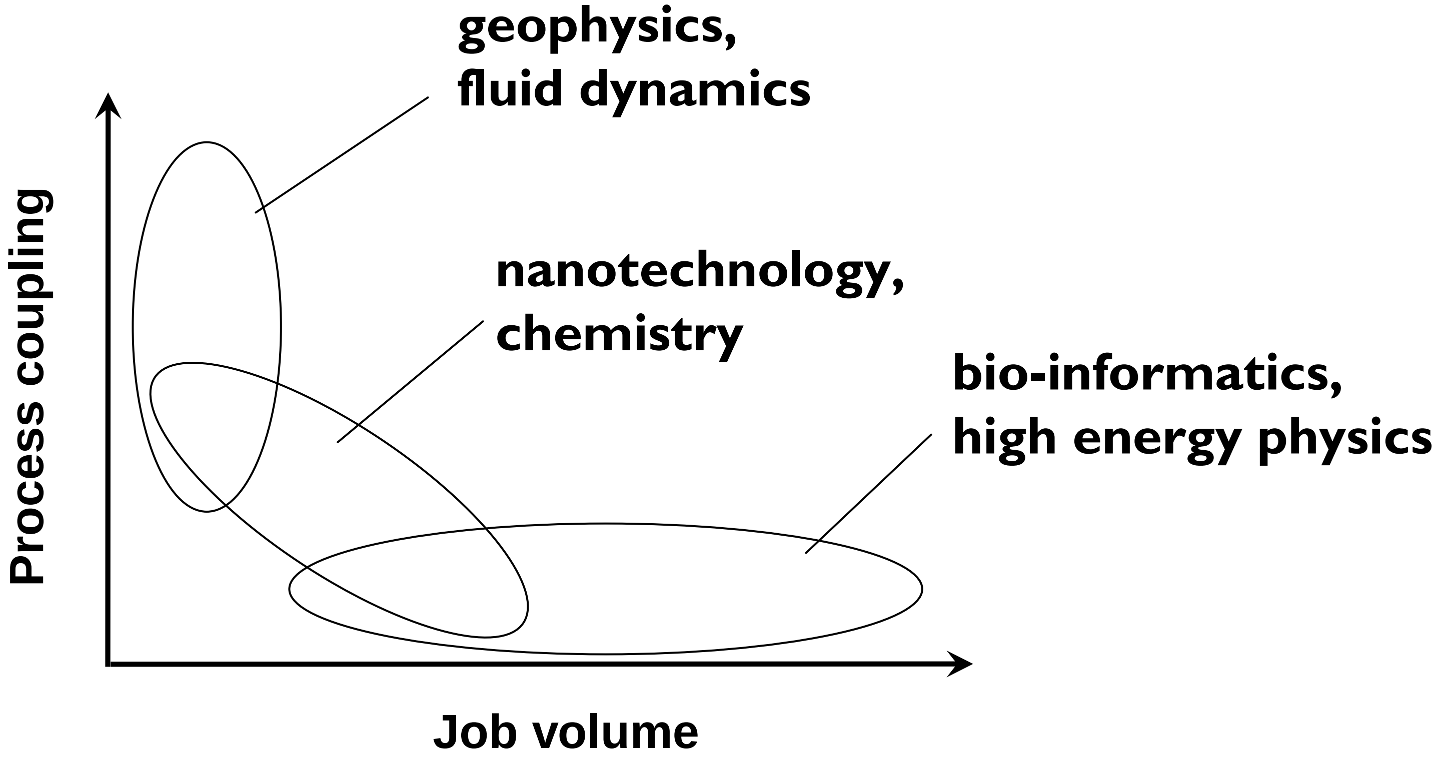
2007	cluster node	4	20 Gflops/s	8 GB	250 GB	0.02 Mkr
2007	cluster	4000	20 Tflops/s	8 TB	250 TB	20 Mkr

Application vs. Hardware



Application vs. Hardware

Application profile:



Usage 2005-2006

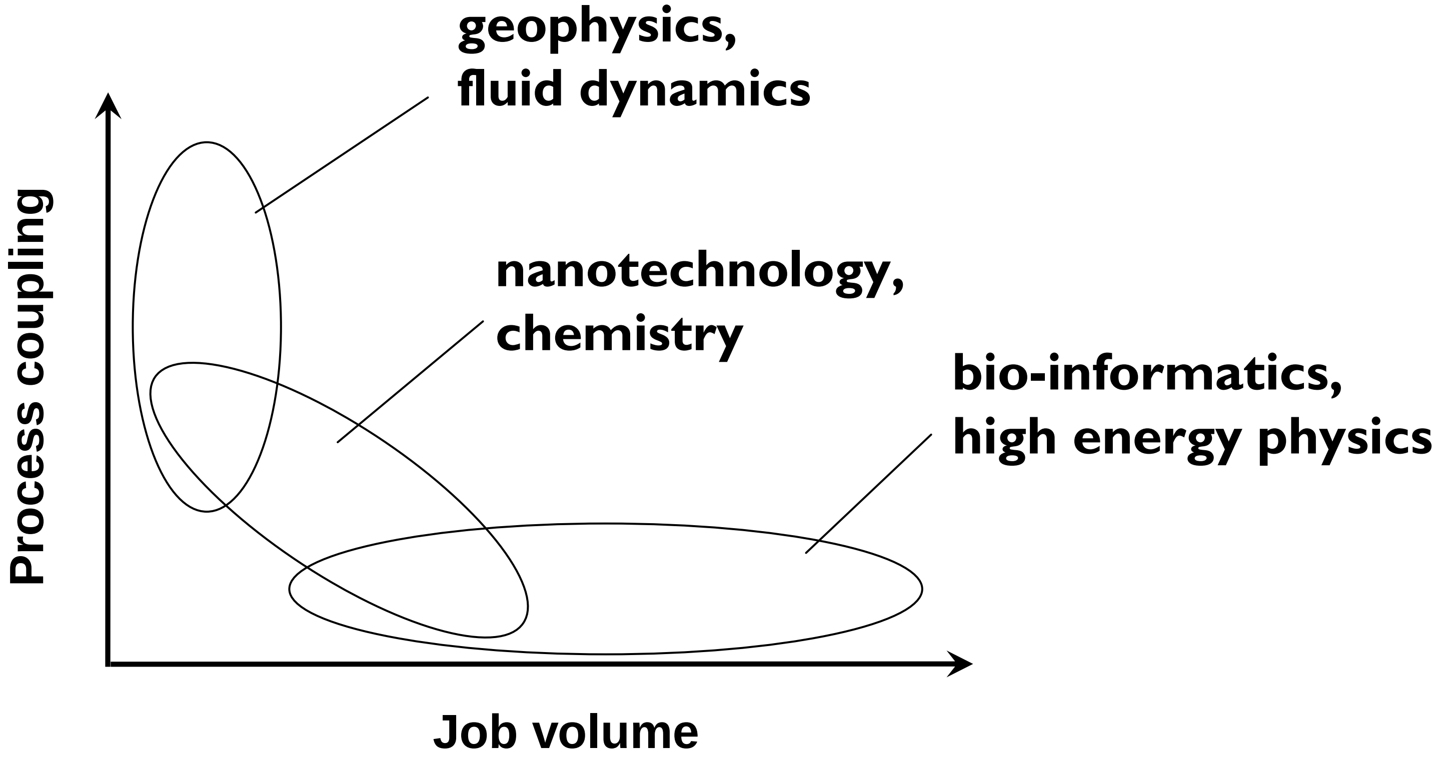
Oct 1, 2005 – Sept. 30, 2006

Organisation		Discipline		Funding	
met.no	24%	geophysics	40%	NORKLIMA	16%
UiO	24%	chemistry	34%	University	16%
UiT	17%	physics	13%	SFF / SFI	10%
UiB	12%	…		EU	9%
NERSC	8%			NANOMAT	8%
NTNU	7%			FRINAT	7%
…				YFF	6%
				…	

Between 55-65 active projects
Ca. 200 TB archived data (mostly climate)

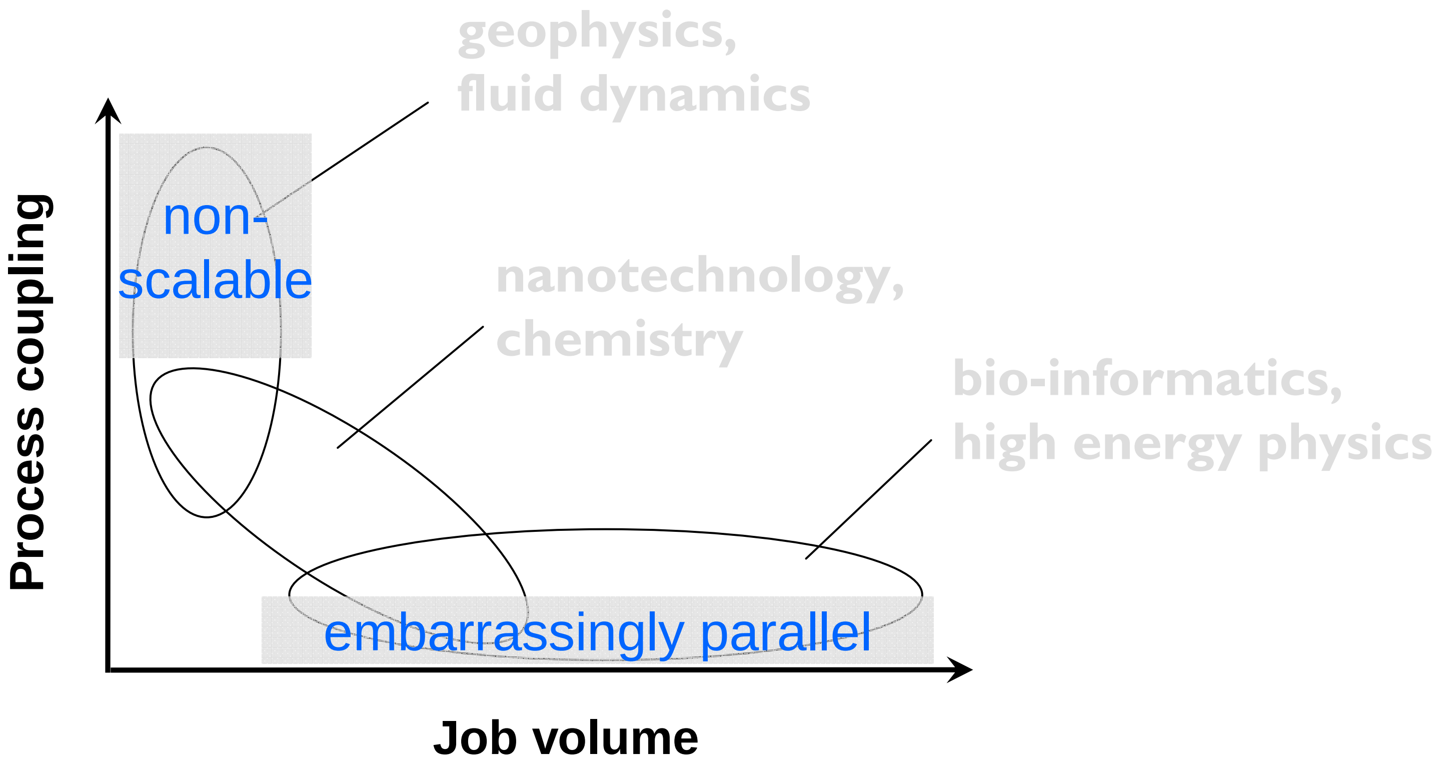
Application vs. Hardware

Application profile:



Application vs. Hardware

Application profile:

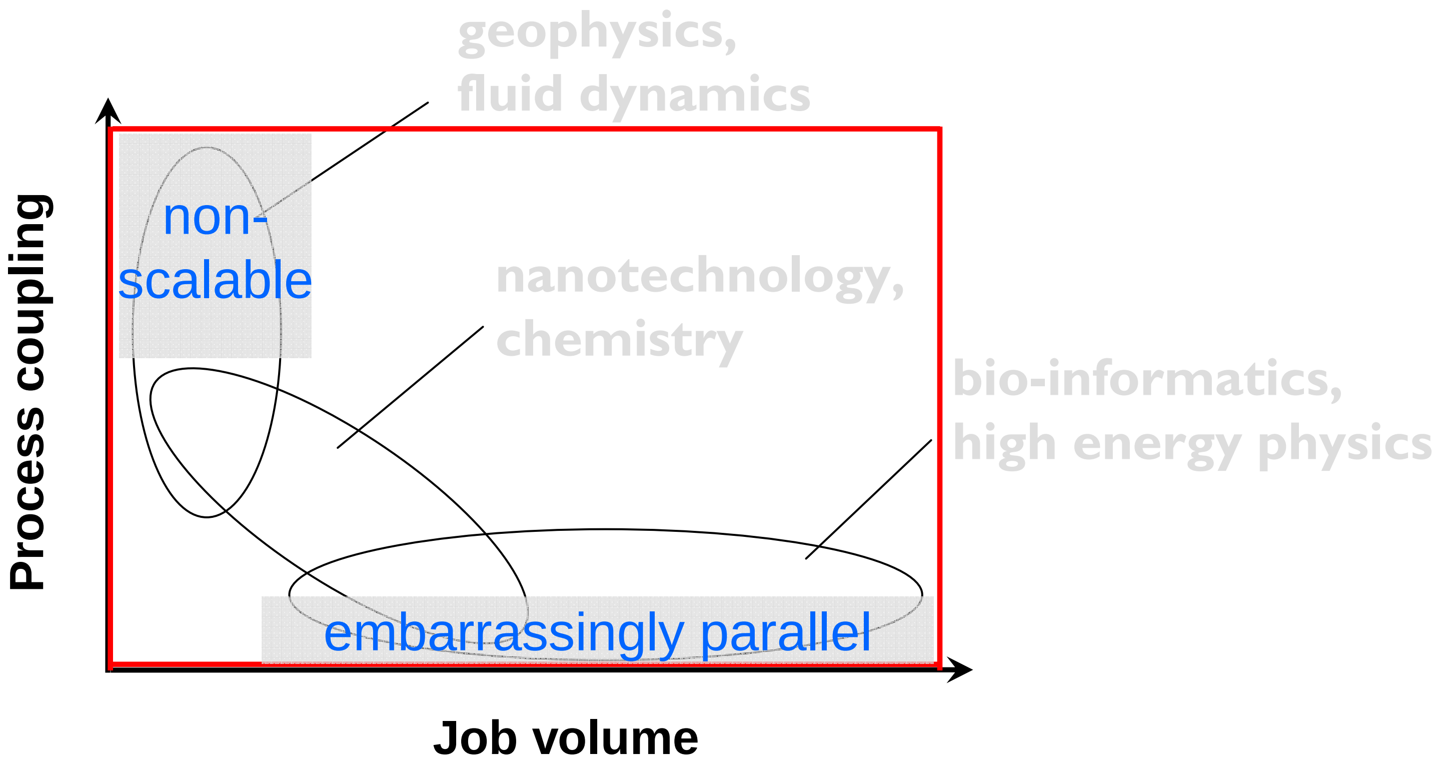


non-scalable : bandwidth / latency bound applications

embarrassingly parallel : many independent tasks

Application vs. Hardware

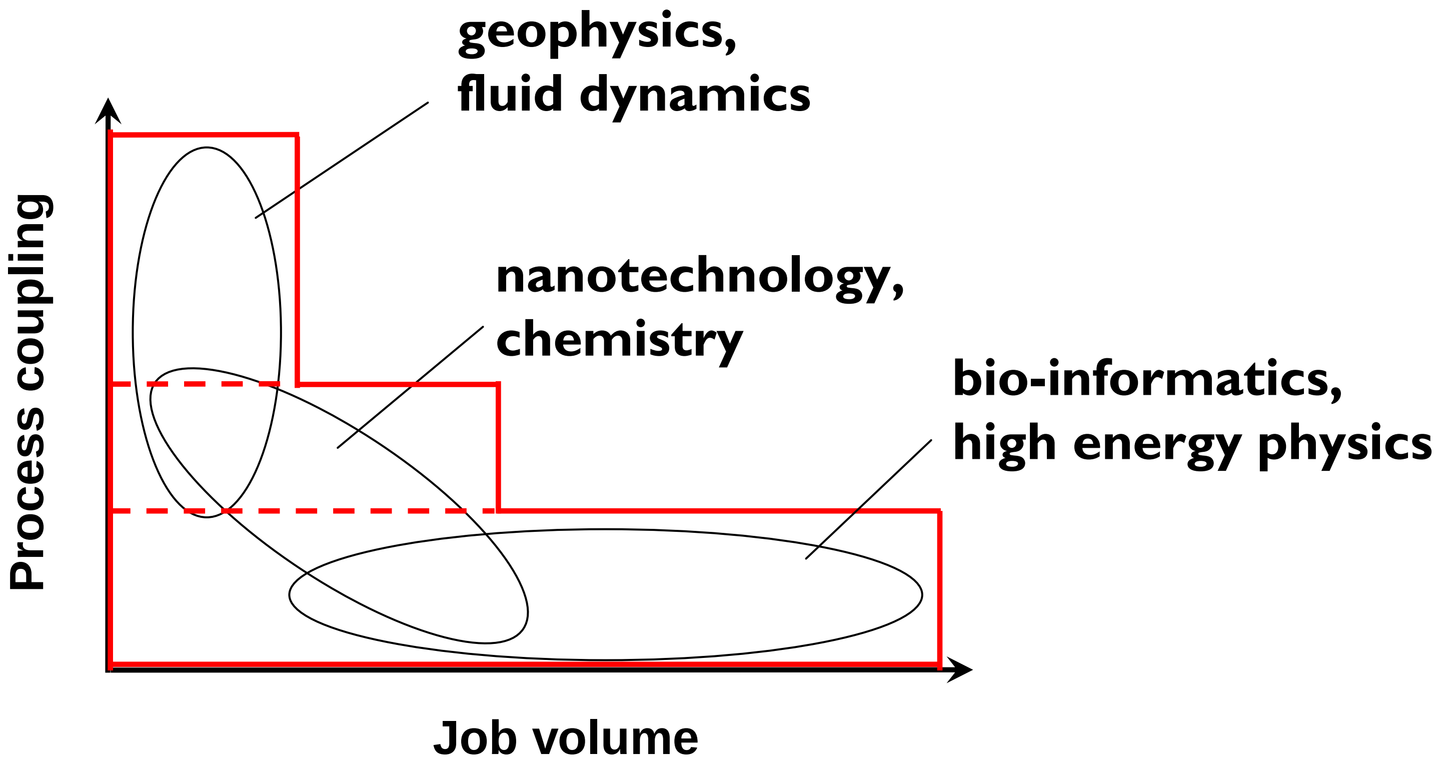
Hardware profile:



: one solution that fits all – not cost-efficient

Application vs. Hardware

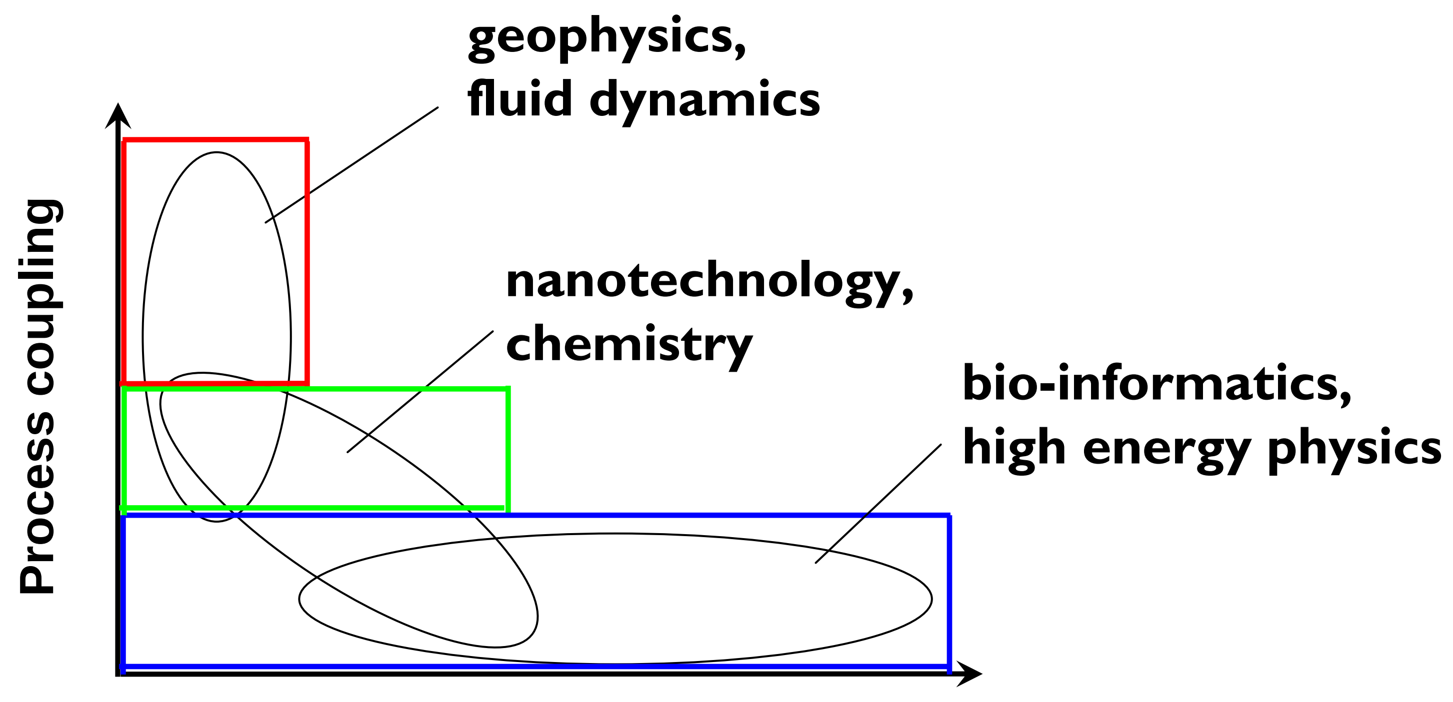
Hardware profile:



: cover what is needed (multiple machines)

Application vs. Hardware

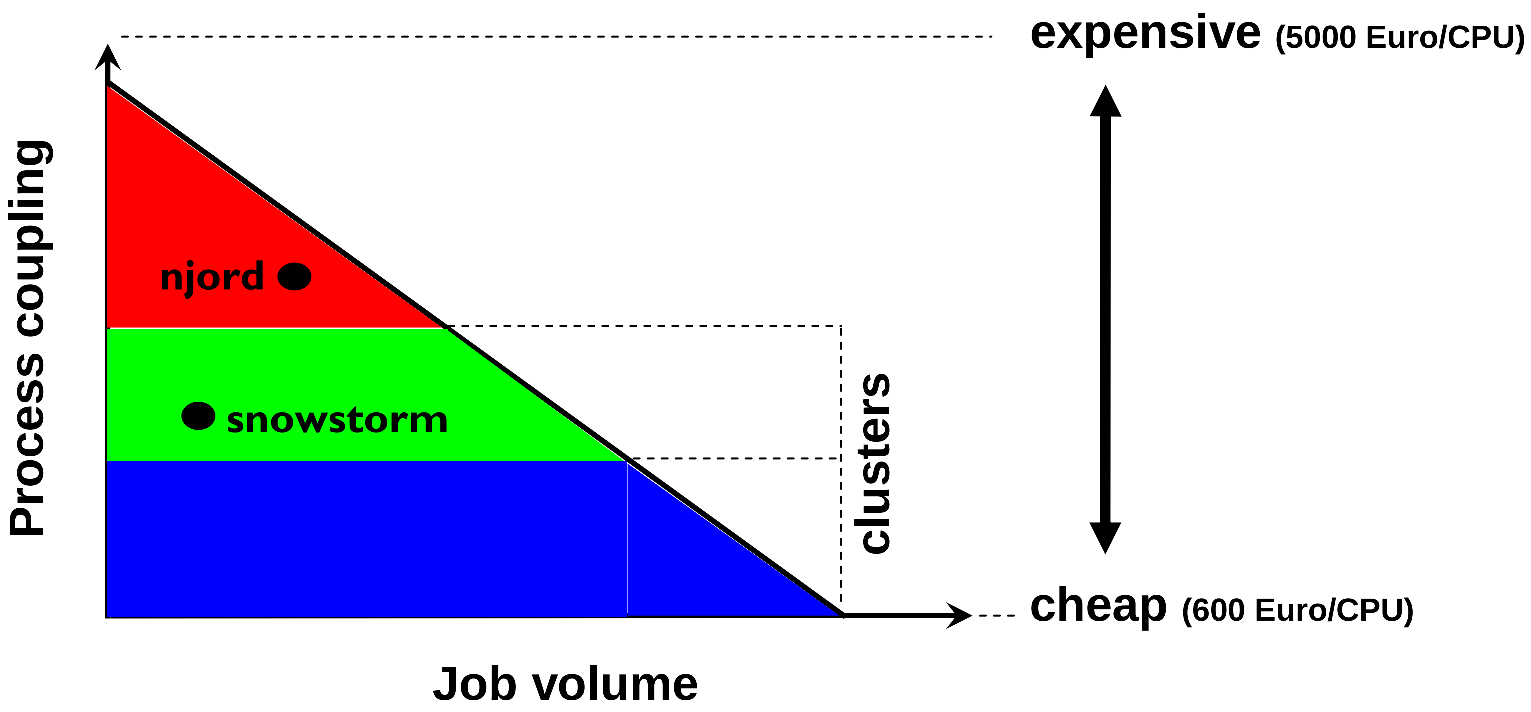
Hardware profile:



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

Application vs. Hardware

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)

NTNU **7 Tflop/s**

njord.hpc.ntnu.no
992 power5+ CPUs, 2 TB memory, 70 TB disk

11/2006

2.3 Tflop/s

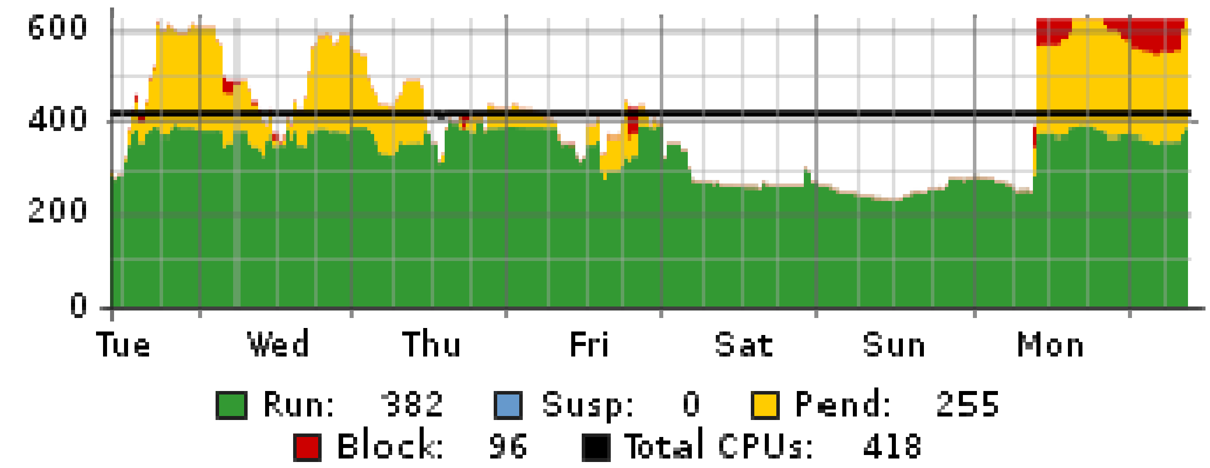
The Snowstorm Cluster

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

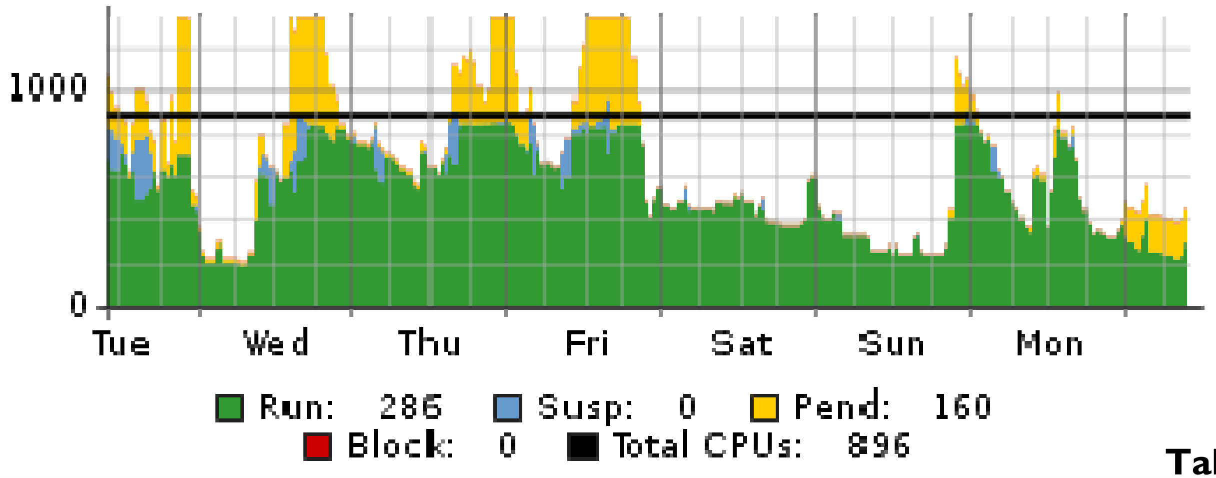
09/2005

Photo © Thilo Bubek – University of Tromsø

Status:



snowstorm



njord

Taken from www.notur.no
Tue Jan 30 2007 9.30 am

hardware

Partner	System	CPU	#CPUs	#Gflops	until
Notur installations					
NTNU	SGI Origin 3800	r14000	896	1000	12/2006
UIB	IBM p690 Regatta	power4	96	499	03/2007
UIB	IBM 1300 cluster	Pentium	64	80	03/2007
UIO	HP SuperDome	Itanium2	64	384	09/2007
UIT	HP rx4640	Itanium2	418	2300	09/2008
NTNU	IBM p575+	power5+	992	7000	11/2010

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

Partner	System	CPU	#CPUs	#Gflops	until
Local clusters					
UiB	IBM e1350 cluster	AMD	172	825	
UiO	Dell cluster	Xeon	384	2200	
NTNU	Dell cluster	Xeon	64		

future hardware

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:

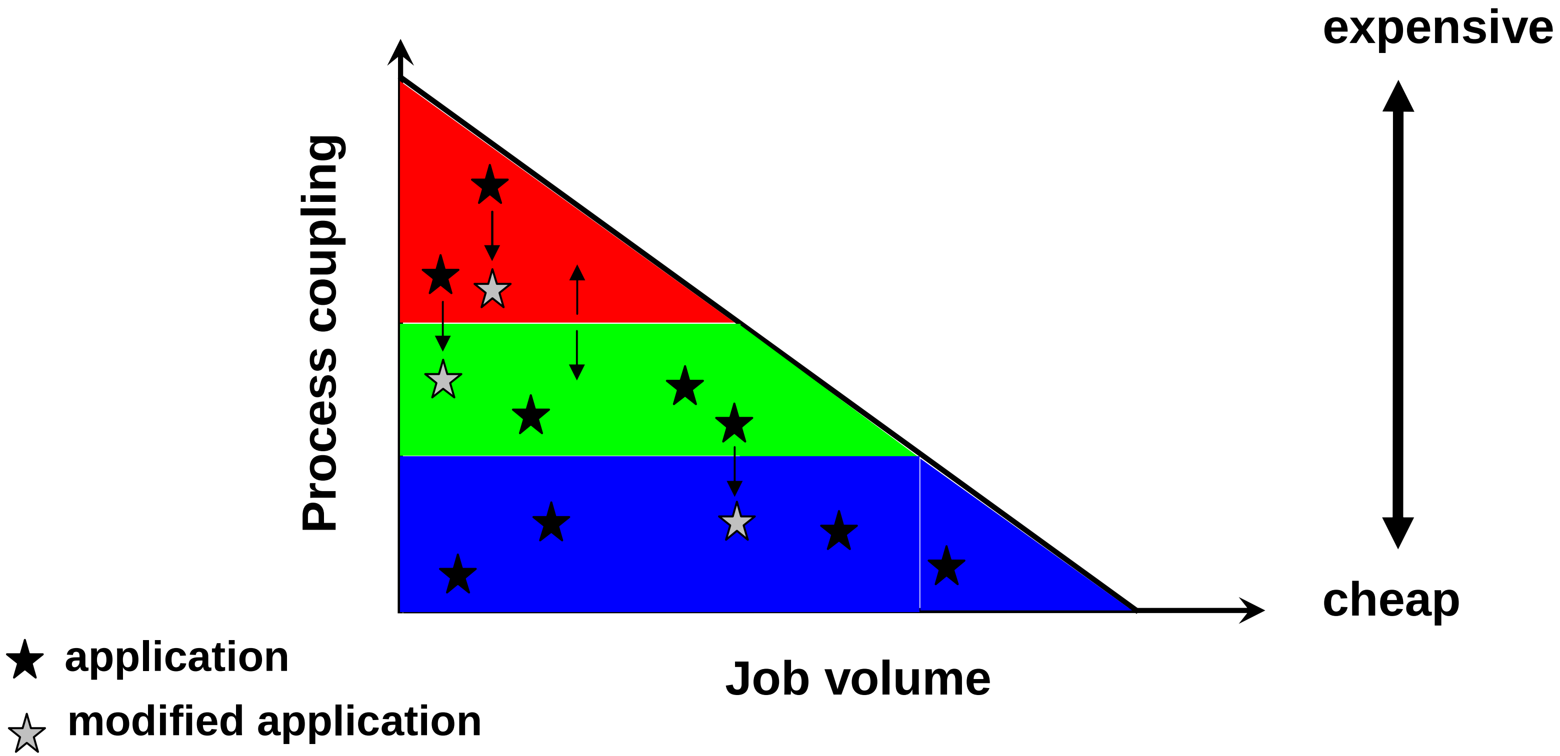
- Heat problems prevent increase in clock-speed : multi-core
- Hardware investments in coming years will have > 1000 cores

Software:

- Problem/Model sizes increase
- Need to improve scalability of software

Application vs. Hardware

Performance pyramid:



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

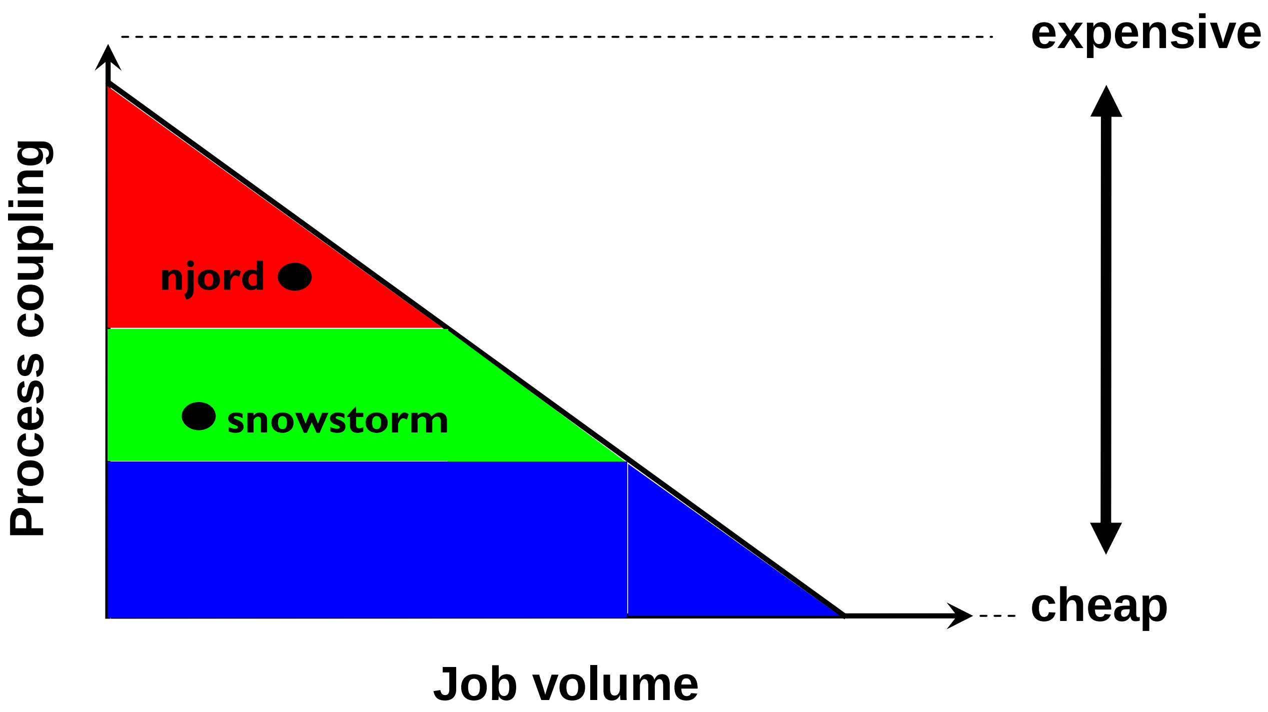
Arvid Lundervold, Department of Biomedicine, UiB
- Quantitative brain imaging in health and disease

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

Application vs. Hardware

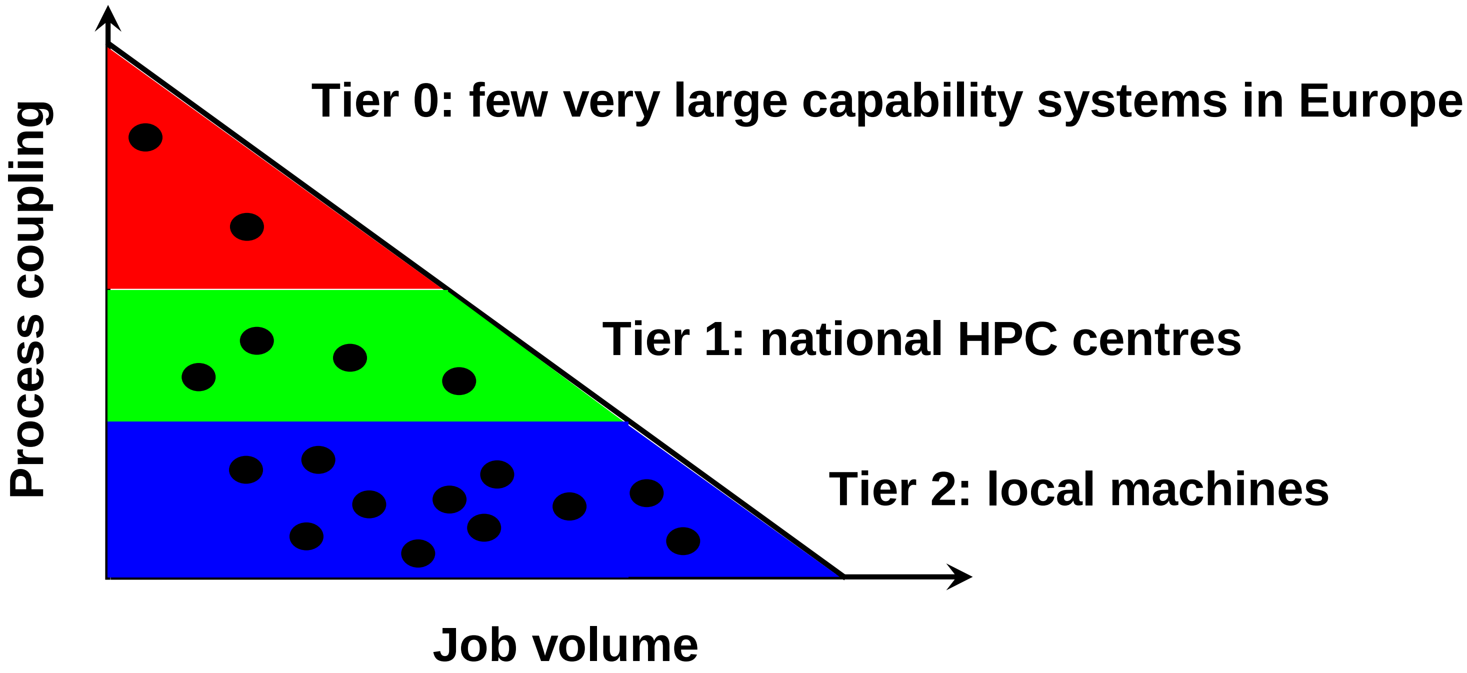
Performance pyramid:



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

Application vs. Hardware

European performance pyramid:



● **machine**

- capacity: clusters with default node interconnect (embarrassingly parallel, MPI)
- capacity: clusters with fast node interconnect (MPI)
- 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)

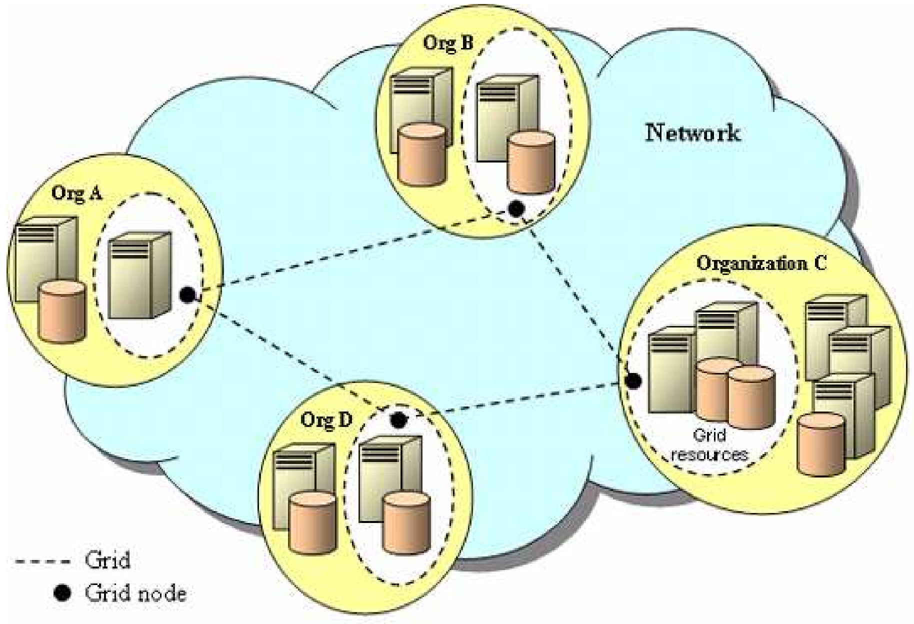
Better attention to storage:
1. Decoupling of data/storage from computer → distributed solutions
2. Optimize location of data (repositories, replication, …)
3. Long-term storage of data (surviving shifts in technology)

Better attention to storage management coincides with upgrade of the national research network from 2.5 Gbit/s to 10 Gbit/s

NorGrid – Norwegian Grid Infrastructure

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, …

NorGrid – Norwegian Grid Infrastructure

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

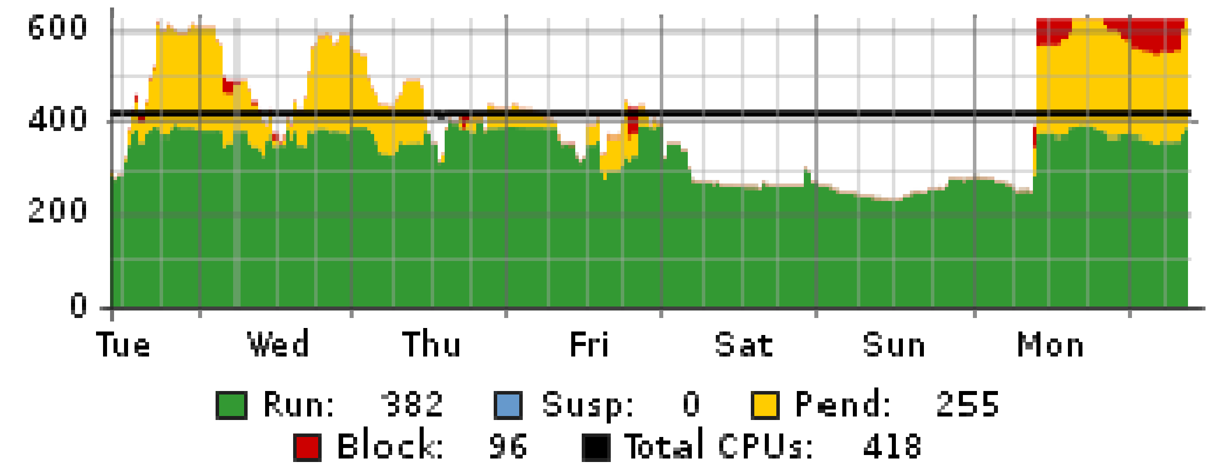
1. Job brokering: user submits job to 'the grid' (with well-defined constraints)

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

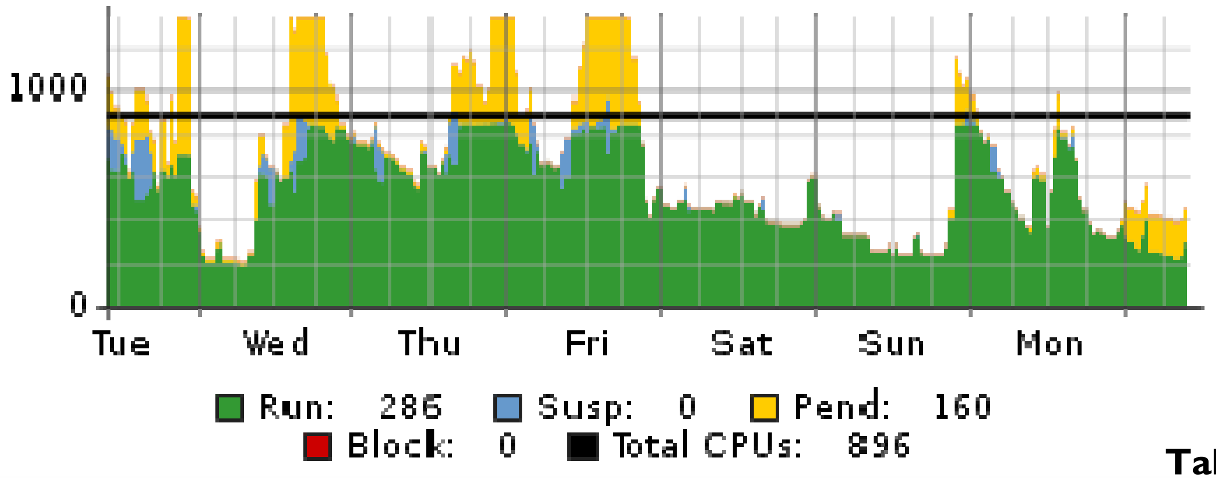
3. Advanced services to collect, gather, share and publish data

NorGrid – Norwegian Grid Infrastructure

Job brokering for resource scheduling:



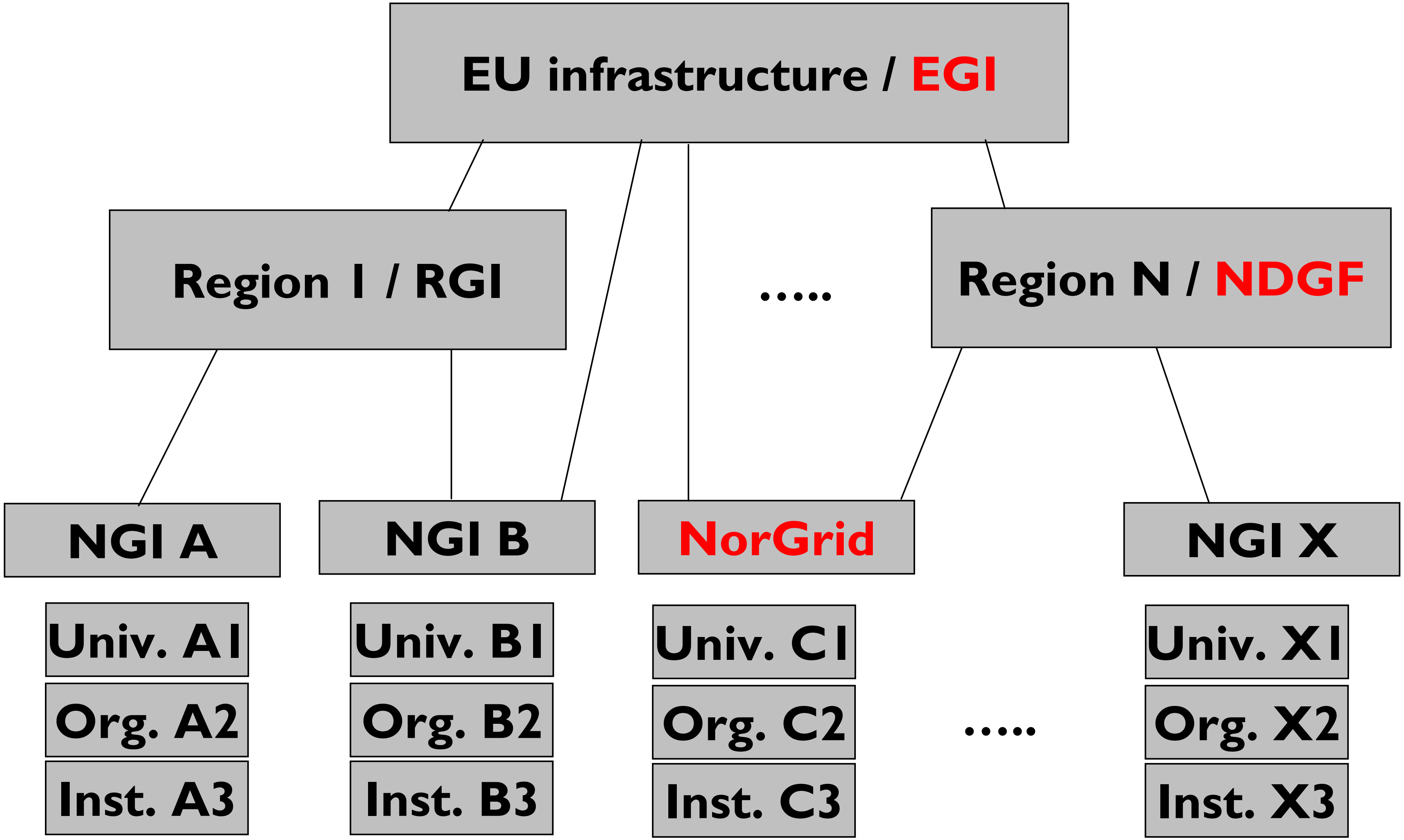
snowstorm



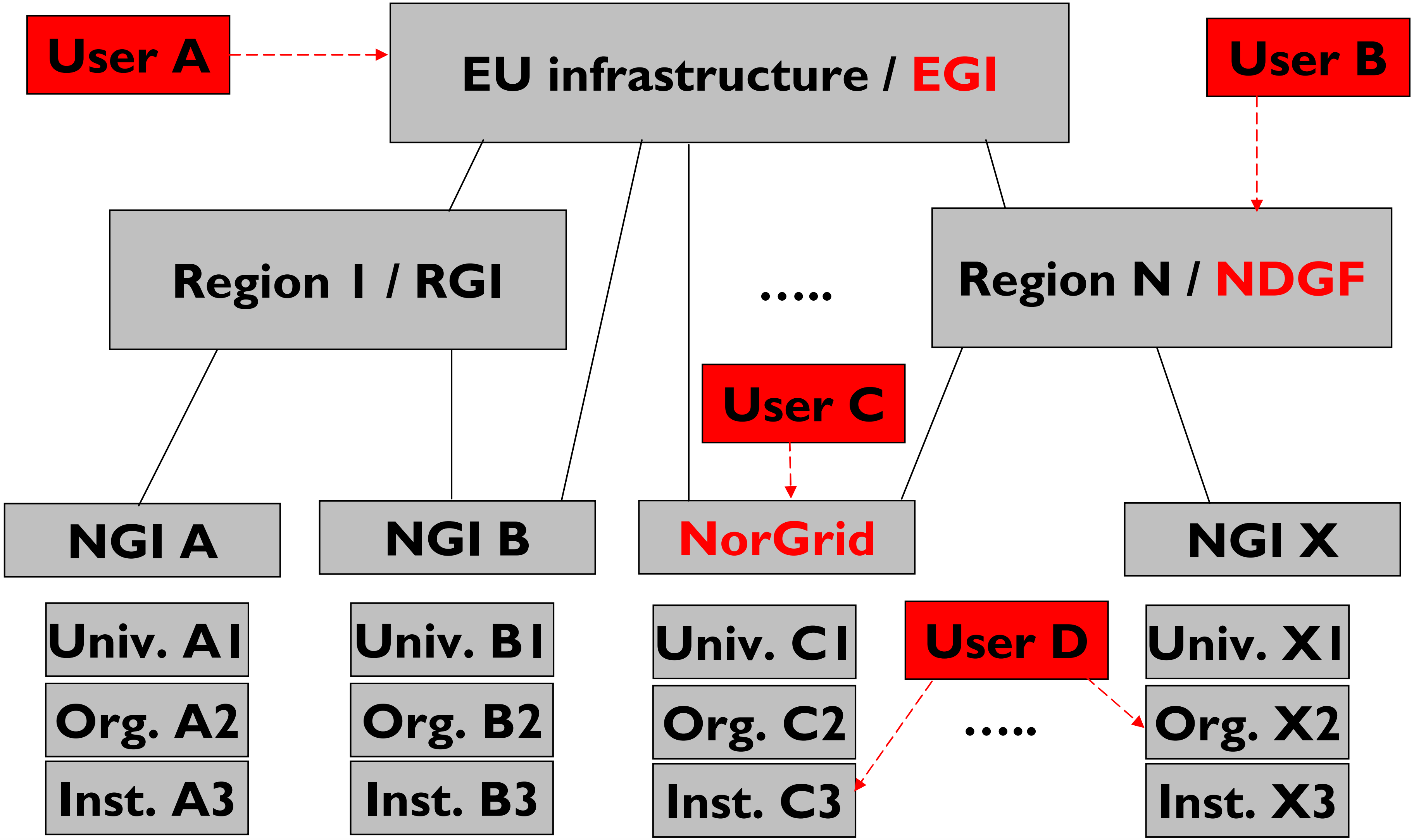
njord

Taken from www.notur.no
Tue Jan 30 2007 9.30 am

EGI - European Grid Infrastructure



EGI - European Grid Infrastructure



Grid application: LHC accelerator

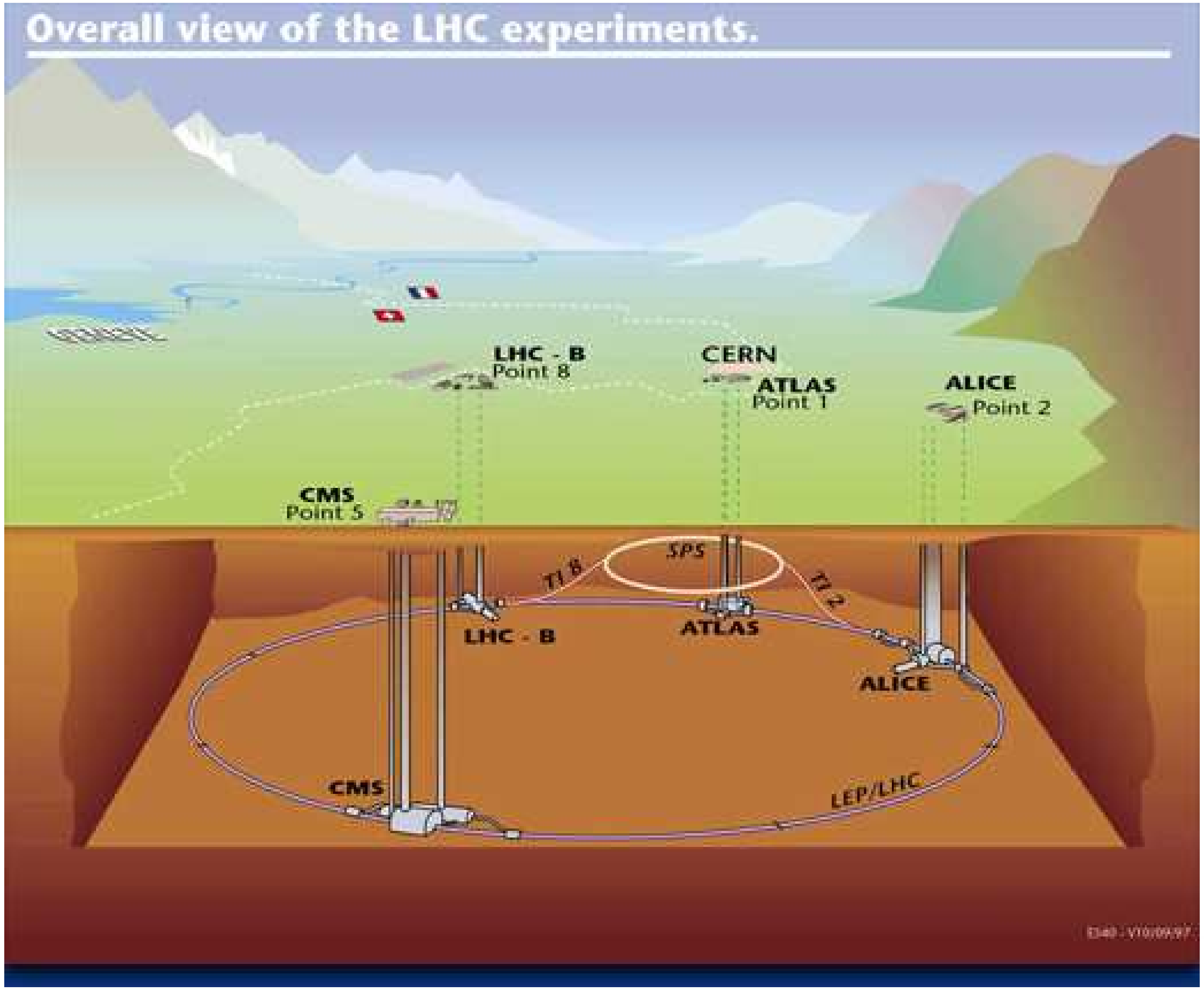
The Large Hadron Collider (LHC) at CERN is a particle accelerator which will probe deeper into matter than ever before. Due to switch on in 2007, it will ultimately collide beams of protons at an energy of 14 TeV. Beams of lead nuclei will be also accelerated, smashing together with a collision energy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

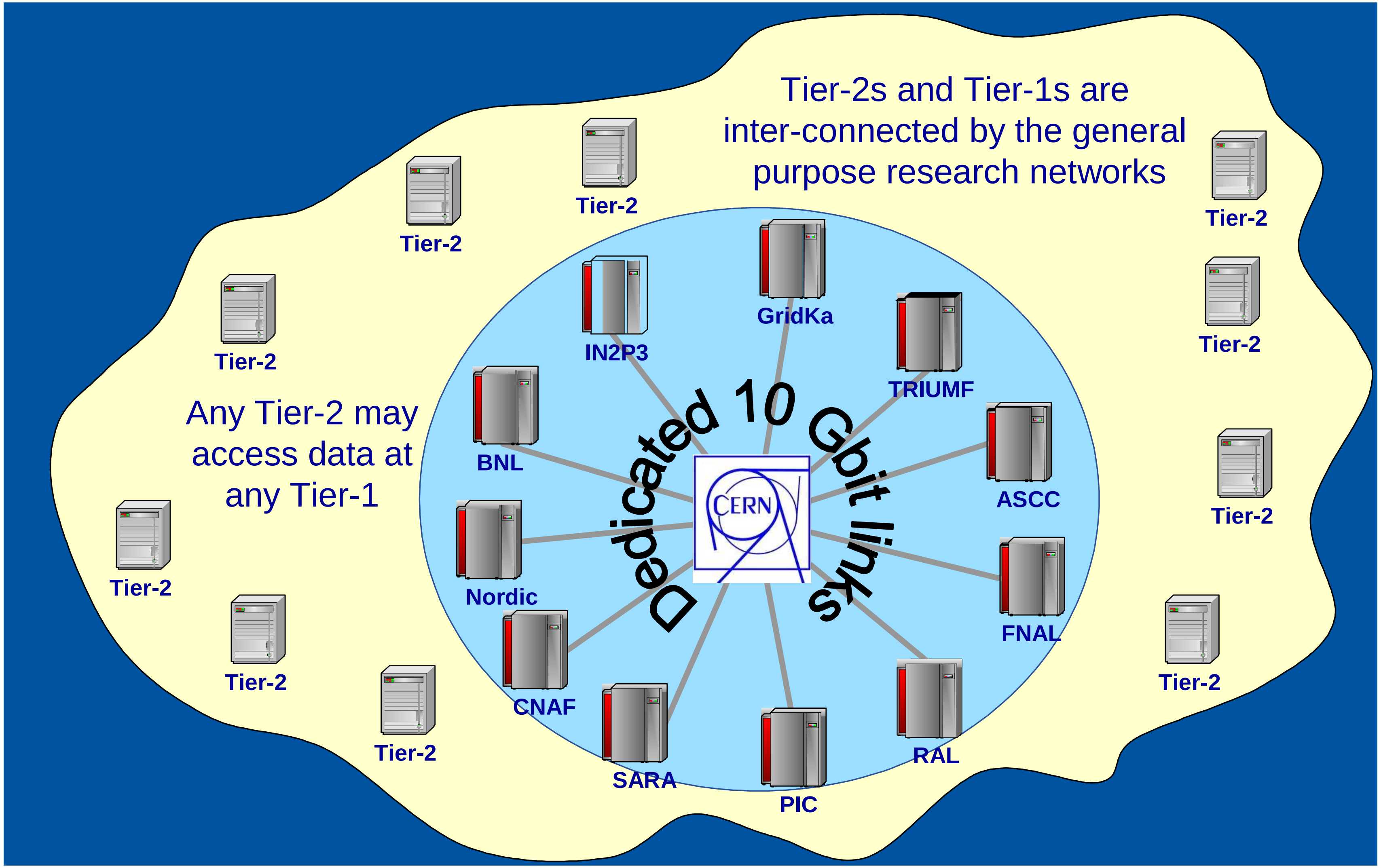
Five experiments: ATLAS, ALICE, CMS, LHCb, TOTEM



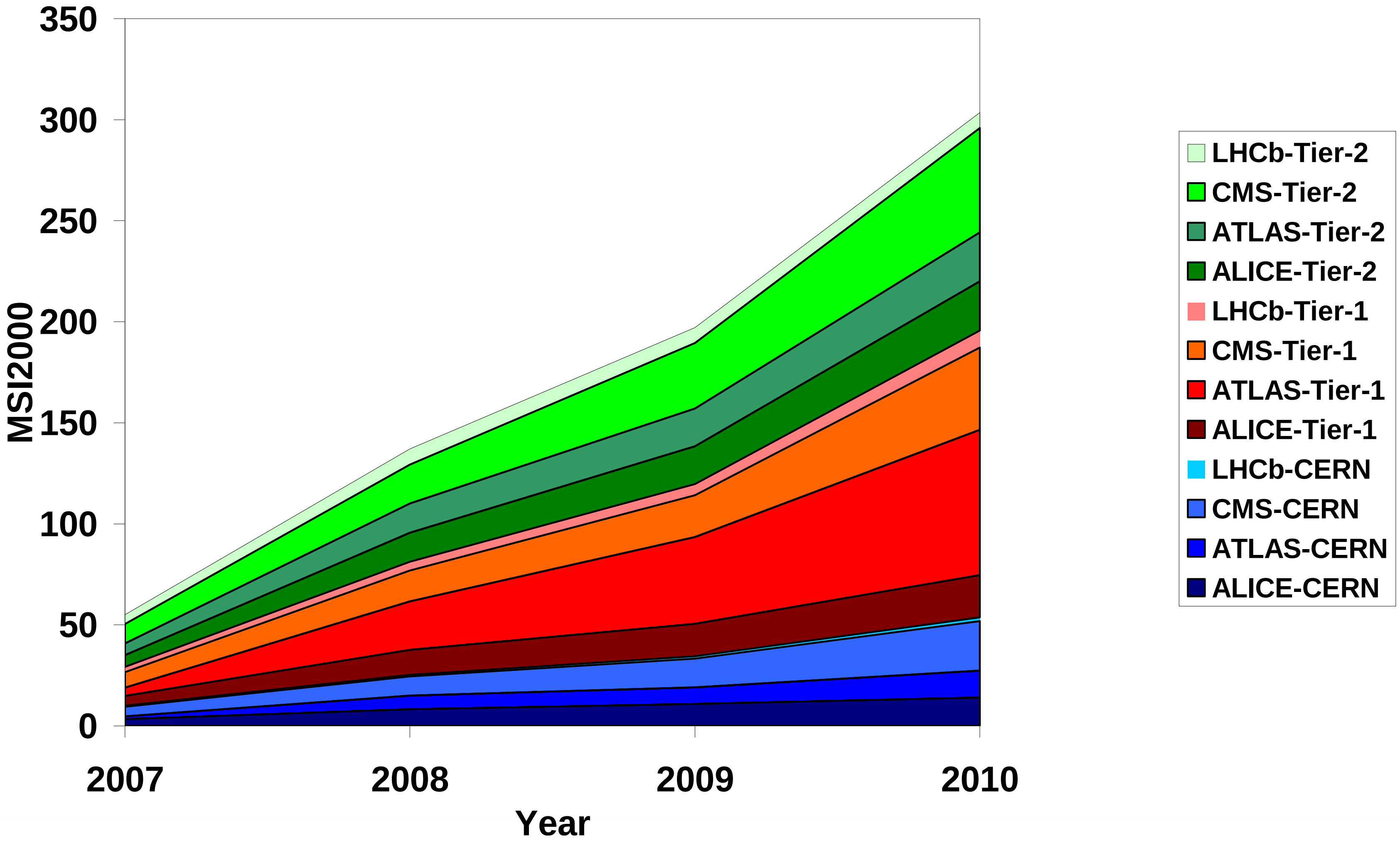
WLCG - Hierarchy

- 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)

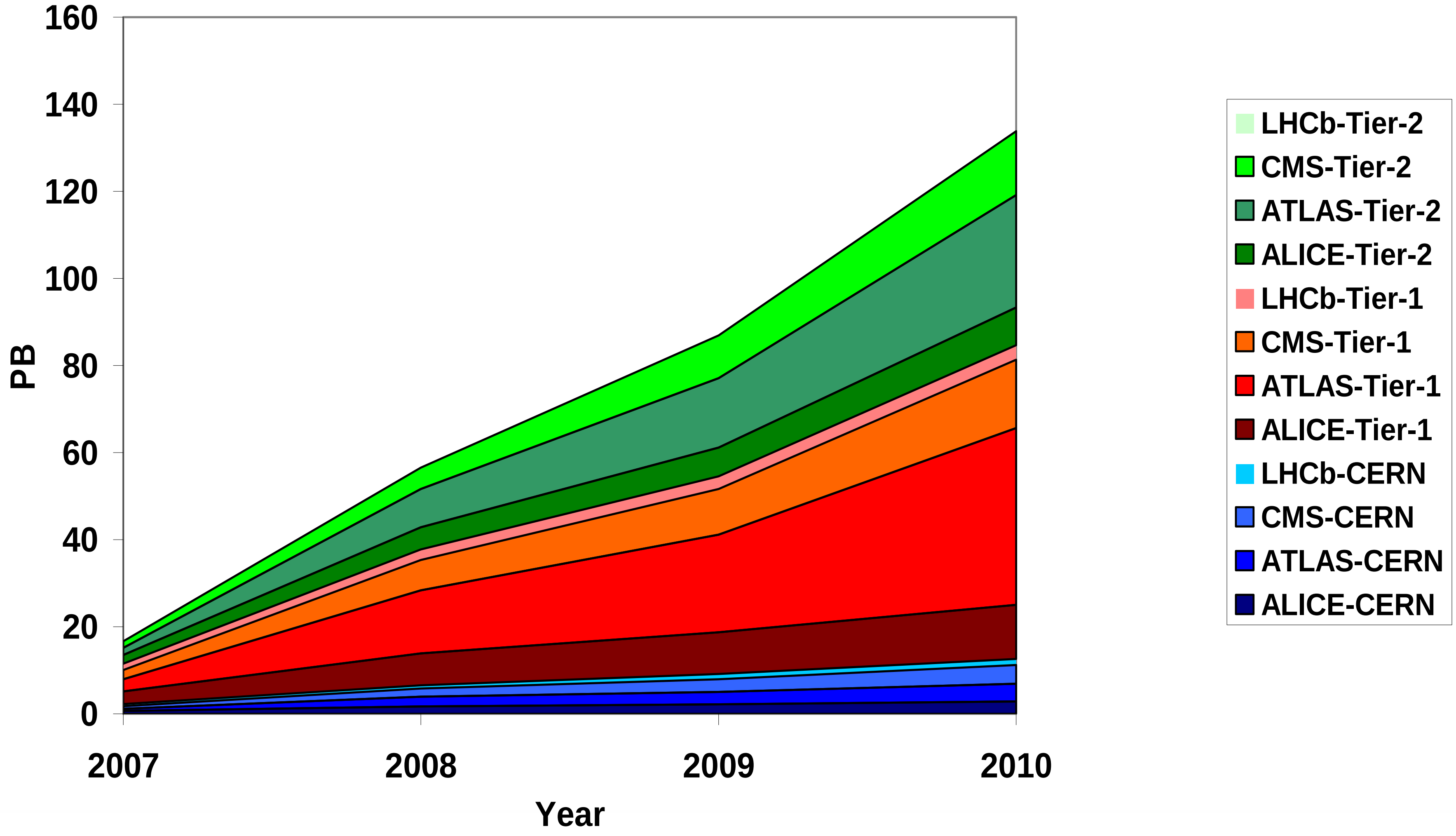
WLCG - Tier-0-1-2 connectivity



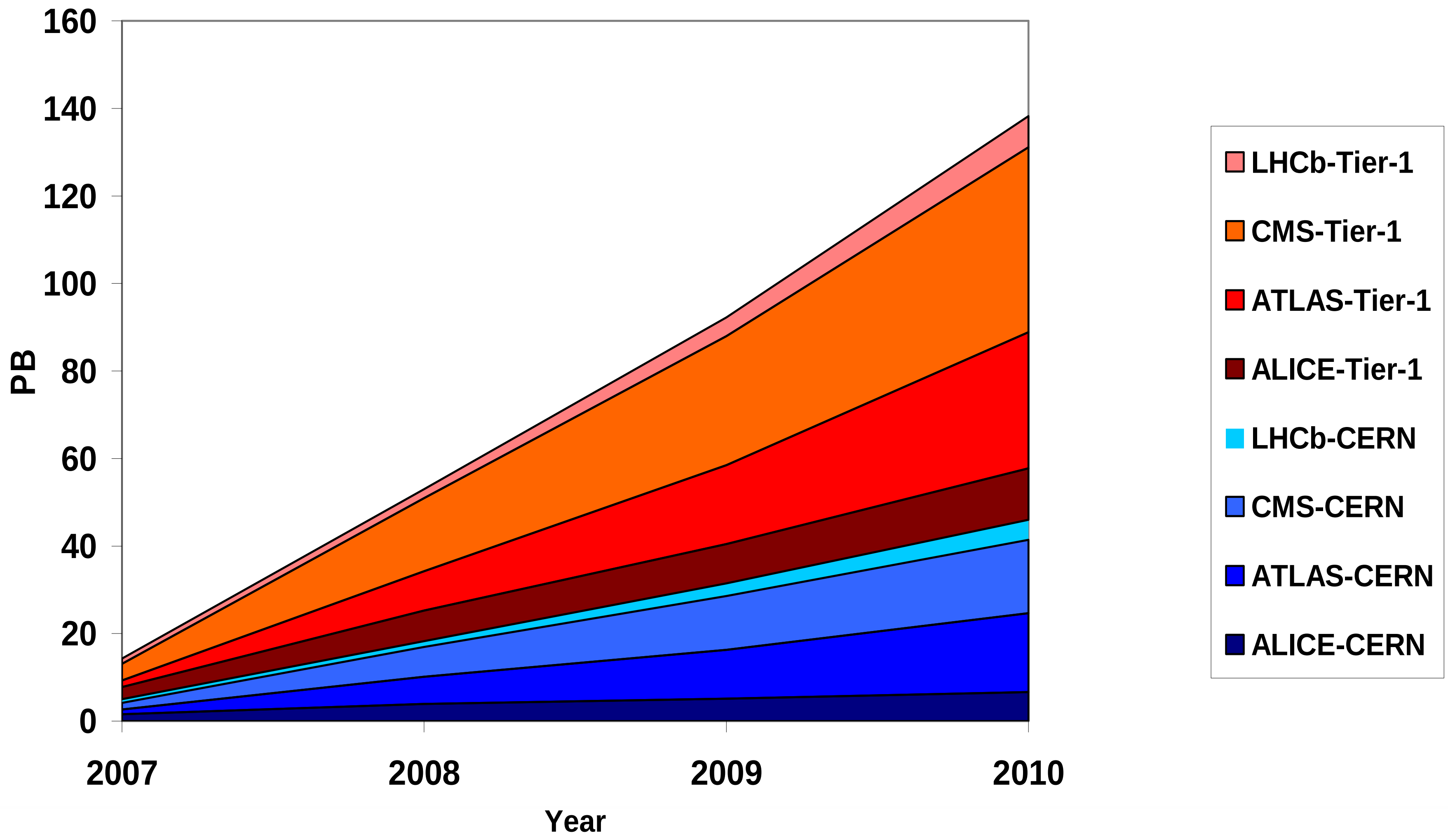
WLCG - CPU requirements



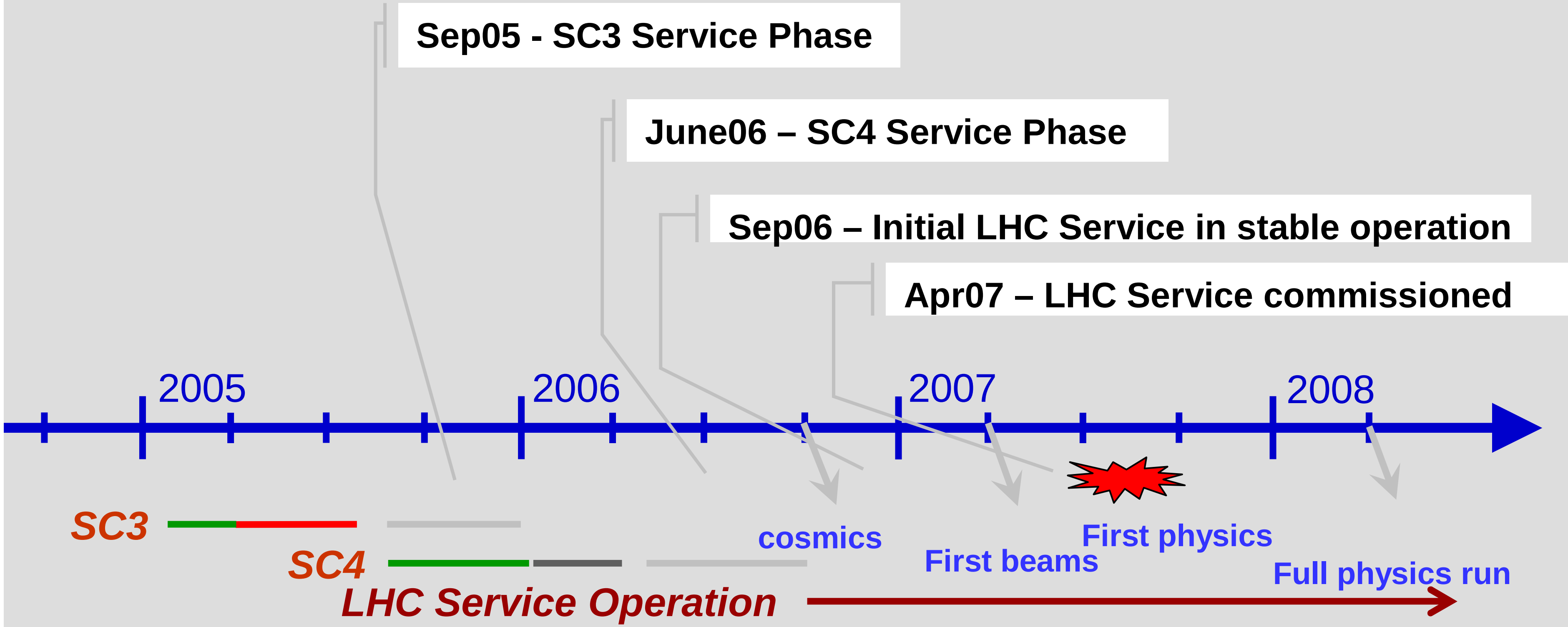
WLCG - Disk requirements



WLCG - Tape requirements



LHC timeline



NDGF

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

Nordic Tier-1 commitments for LHC (ALICE/ATLAS/CMS)

	2006	**2007**	**2008**	**2009**	**2010**
CPU (kSI2K)	520	1340	2610	3470	4280
Disk (Tbytes)	160	440	890	1500	2200
Tape (Tbytes)	240	435	872	1500	2260
WAN (Mbits/sec)	2000	5000	10000	20000	20000

Nordic Tier-1

Norwegian contribution is derived from 20% of Nordic Tier-1:

	2006	2007	2008	2009	2010
CPU (kSI2K)	264	528	694		
Disk (Tbytes)	75	200	325		
Tape (Tbytes)	60	160	260		
WAN (Mbits/sec)	2500	10000			

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