TSEA28, Datorteknik Y
Guest Lecture
15 May 2013

# NSC

## National Supercomputer Centre

Network · Storage · Computing

Niclas Andersson, nican@nsc.liu.se
National Supercomputer Centre, Sweden

# Contents

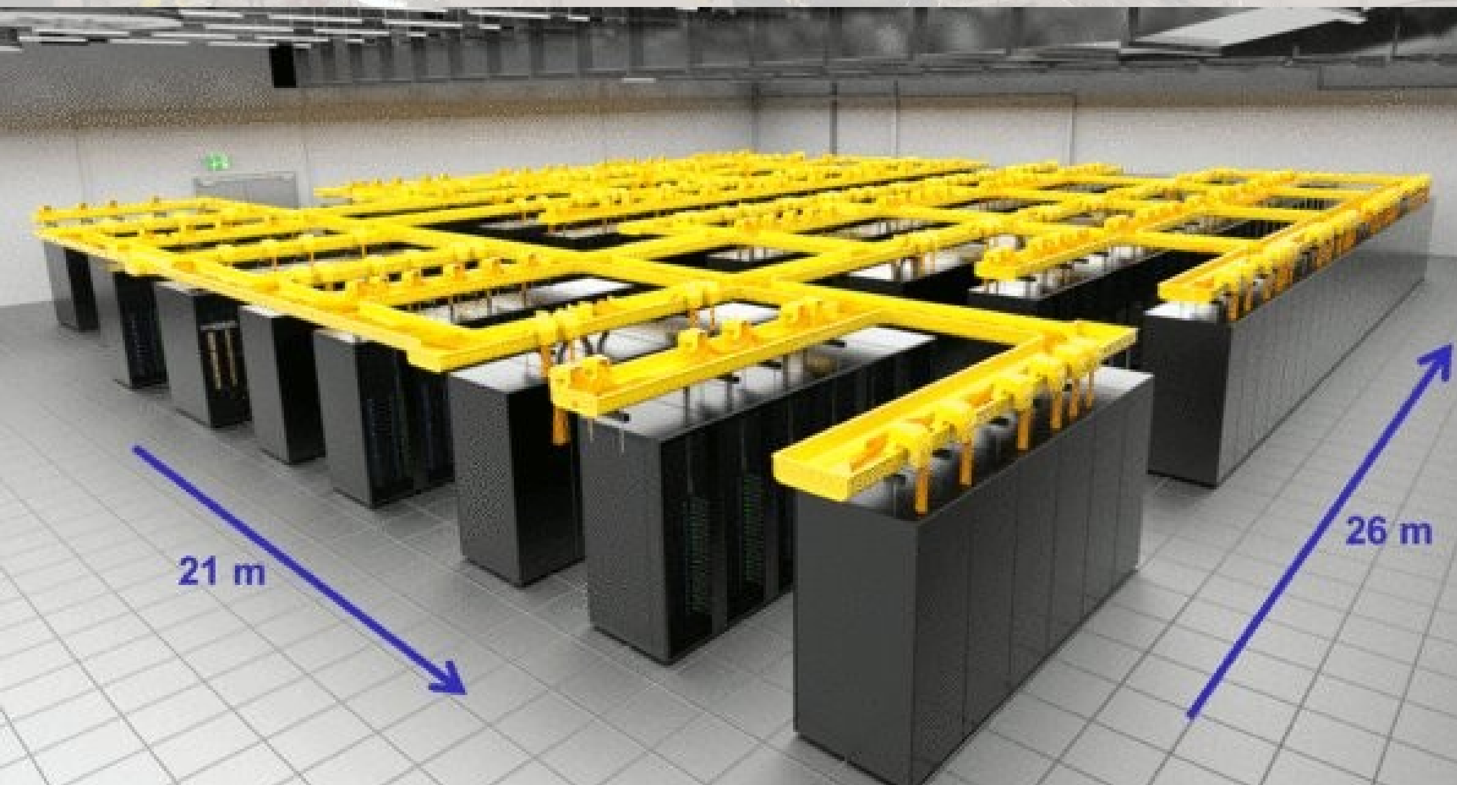
- High Performance Computing
- Processors of today
  - examples: Intel, AMD, NVIDIA, ...
- National Supercomputer Centre
- Large scale computing resources
- Applications

Slides: http://www.nsc.liu.se/~nican/education/tsea28_2013.pdf

# What is a Supercomputer?

**Cray-1A**

# What are the differences?




#6 on top500: SuperMUC (#1 in Europe)


# What are the similarities?

# The most important aspects for High Performance Computing (HPC)

- Floating point operations per second
- Memory bandwidth
- Interconnect performance (bandwidth, latency)
- Parallelism, parallelism, parallelism
- Power consumption
- Efficient algorithms and good programming

# Parallelism, parallelism, parallelism

In core

- Many ALUs
- Pipelining
- Vectors; SSE, AVX
- Instructions: FMA, ...
- Out-of-order execution
  - Shadow registers
  - Speculative execution
- Hyper threading (Intel)

On chip

- Many cores
- Multi level, multi port caches

In server

- Many sockets
- Memory channels
- Co-processors

In system

- Many servers
- Fast interconnect, Infiniband

On site

- Many systems
- Secondary storage

On larger scale

- Collaborative networks
- Grid, Cloud, ...

# Example

Triolith with SandyBridge (Intel Xeon E5-2660)

- 2.2 GHz clock
  (Turbo 3.0 GHz)
- 8 Flop / clock / core
- 8 core / socket
- 2 socket / server
- 1200 compute servers
- 2.2 * 8 * 8 * 2 * 1200 = 338 Tflop/s

# Hybrid computing

Merge traditional CPU with high performance co-processor

Examples:

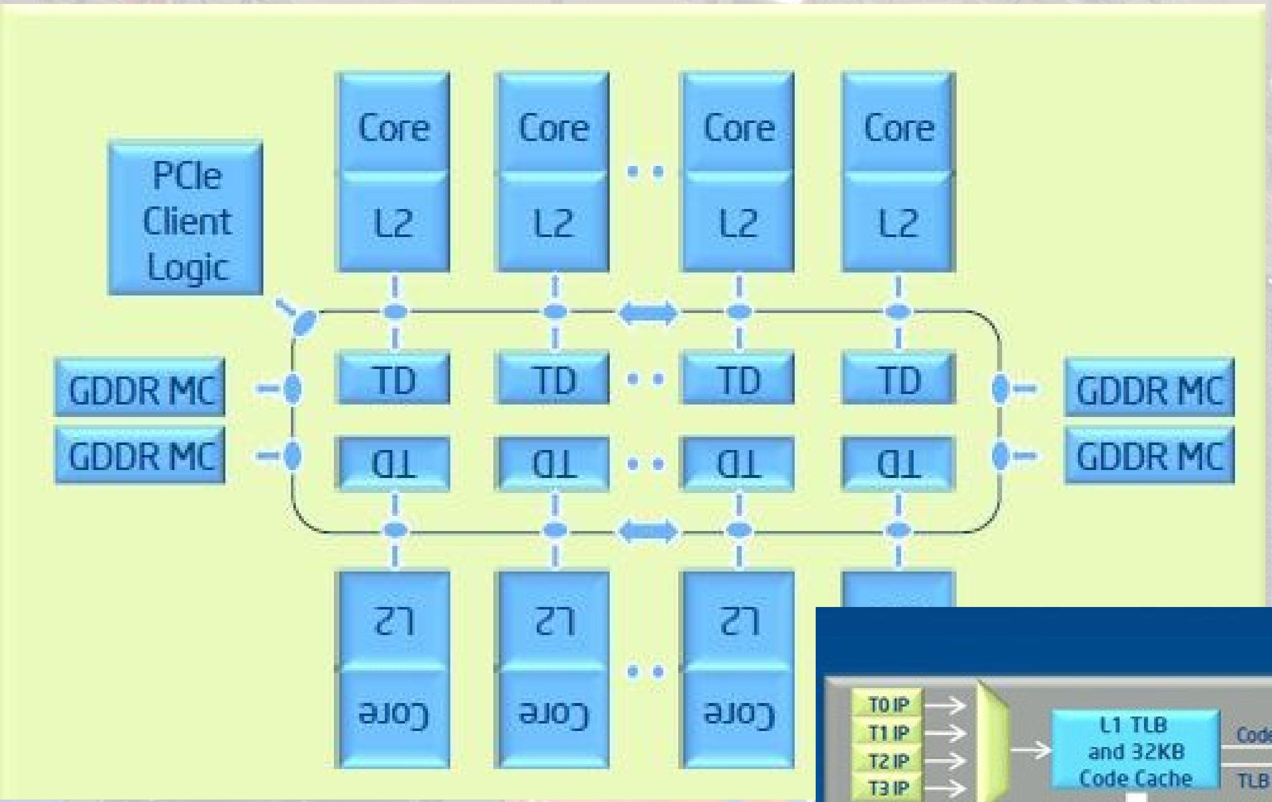
- NVIDIA Kepler
  - General-purpose computing on graphics processing units (GPGPU)
  - Open Computing Language (OpenCL)
  - NVIDIA CUDA
- Intel Xeon Phi
  - Accelerator
  - Intel Compilers & Tools

# NVIDIA Kepler (GK110)

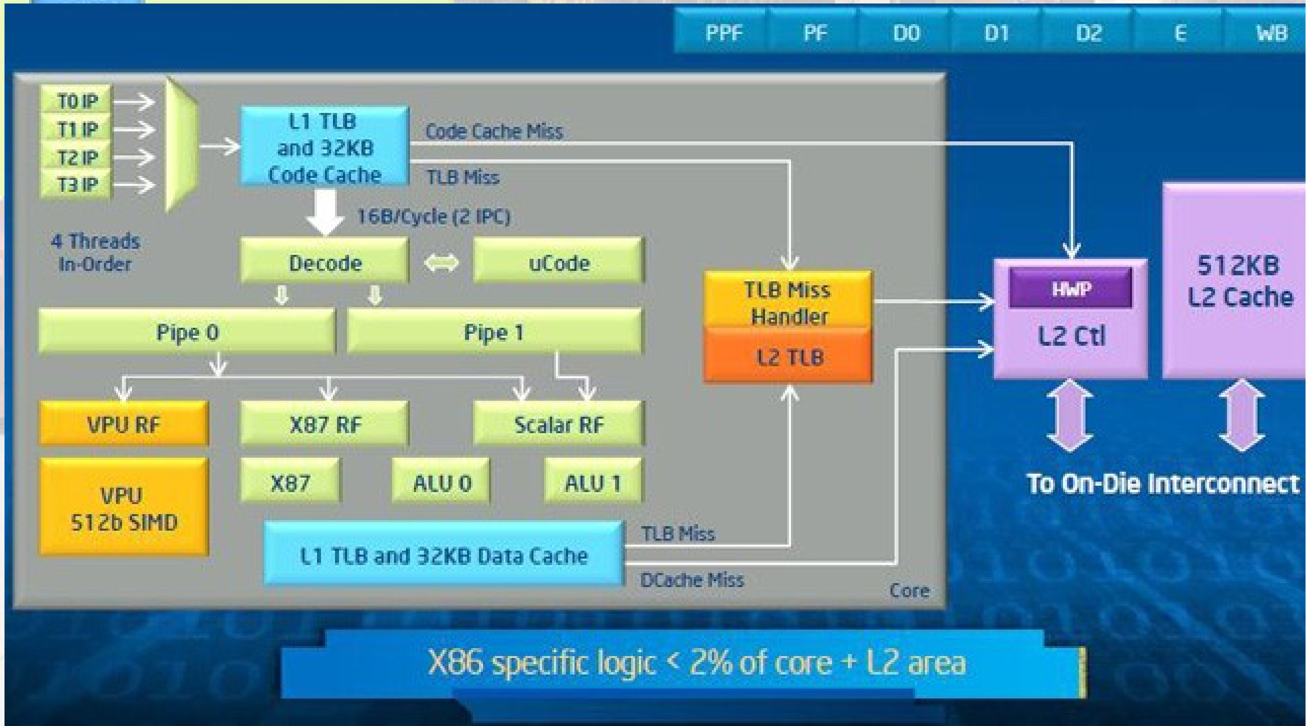
# NVIDIA GK110 SMX Streaming Multiprocessor

- 192 FP32/INT
- 64 FP64
- Kepler: 3 x perf/watt to Fermi
- Dynamic parallelism
- GPUdirect

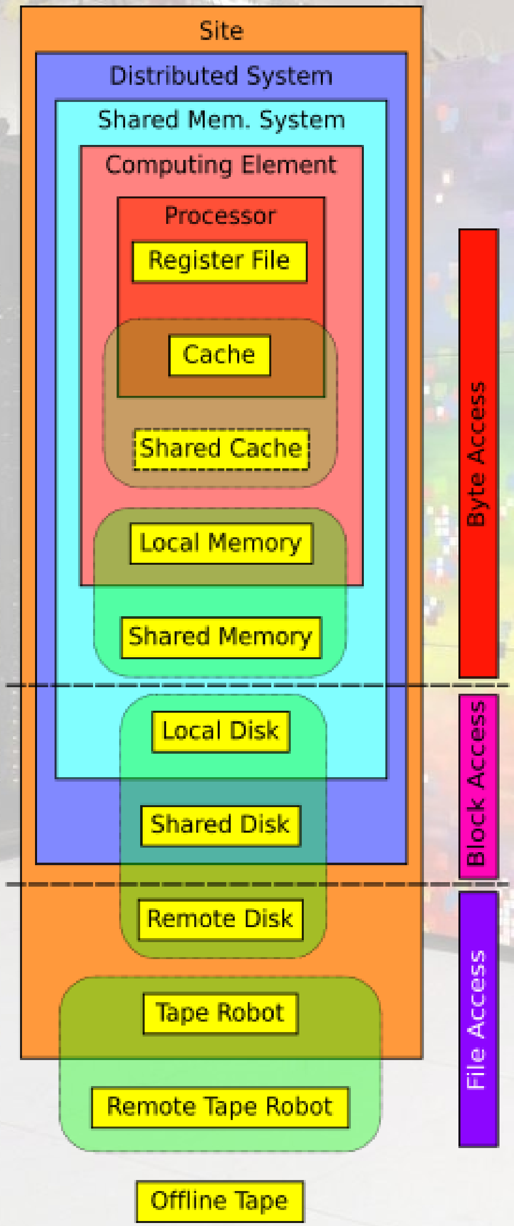
# Intel Xeon Phi




>50 cores



# Storage hierachy




Moving data is expensive

- Distance from ALU
- Performance (bandwidth & latency)
- Size
- Cost (investment & energy)

# Bandwidth vs. Latency

**SNAP** – SNAil based data transfer Protocol (2005)

- Payload/packet: 4.7 GB
- Parallel protocol: 2 packets/transfer
- Faster than ADSL on short distance
- Outperforms IP over avian carriers (1999)

Never underestimate the bandwith of a truckload of tapes on a highway!

# Efficient Algorithms

- Utilize available parallelism in the problem
- Adaptive
- Balance load statically and/or dynamically
- Latency tolerant
- Scalable

Amdahl's Law

$$\boxed{S_p = \frac{1}{f + \frac{1-f}{p}}}$$

$S_p$ speedup
$f$ Sequential fraction
$p$ Number of processors

# Programming

- Fortran (most common), C, C++
- Message Passing interface (MPI)

```
#include <stdio.h>
#include "mpi.h"

int main( argc, argv )
int  argc;
char **argv;
{
    int rank, size;
    MPI_Init( &argc, &argv );
    MPI_Comm_size( MPI_COMM_WORLD, &size );
    MPI_Comm_rank( MPI_COMM_WORLD, &rank );
    printf( "Hello world from process %d of %d\n", rank, size );
    MPI_Finalize();
    return 0;
}
```

```
% mpicc -o helloworld helloworld.c
% mpirun -np 4 helloworld
Hello world from process 0 of 4
Hello world from process 3 of 4
Hello world from process 1 of 4
Hello world from process 2 of 4
%
```

# More MPI: sending in a ring

```
#include <stdio.h>
#include "mpi.h"

int main( argc, argv )
int argc;
char **argv;
{
    int rank, value, size;
    MPI_Status status;

    MPI_Init( &argc, &argv );

    MPI_Comm_rank( MPI_COMM_WORLD, &rank );
    MPI_Comm_size( MPI_COMM_WORLD, &size );
    do {
        if (rank == 0) {
            scanf( "%d", &value );
            MPI_Send( &value, 1, MPI_INT, rank + 1, 0, MPI_COMM_WORLD );
        }
        else {
            MPI_Recv( &value, 1, MPI_INT, rank - 1, 0, MPI_COMM_WORLD,
                      &status );
            if (rank < size - 1)
                MPI_Send( &value, 1, MPI_INT, rank + 1, 0, MPI_COMM_WORLD );
        }
        printf( "Process %d got %d\n", rank, value );
    } while (value >= 0);

    MPI_Finalize( );
    return 0;
}
```

```
% mpicc -o ring ring.c
% mpirun -np 4 ring
10
Process 0 got 10
22
Process 0 got 22
-1
Process 0 got -1
Process 3 got 10
Process 3 got 22
Process 3 got -1
Process 2 got 10
Process 2 got 22
Process 2 got -1
Process 1 got 10
Process 1 got 22
Process 1 got -1
%
```

# MPI primitives

The Base:

MPI_Init

MPI_Finalize

MPI_Comm_size

MPI_Comm_rank

MPI_Send

MPI_Recv

Communication modes:
Blocking, Non-blocking, Buffered, Synchronous, Ready

Collective communication

Group and communicator management

Derived datatypes

Virtual topologies

One-sided communication

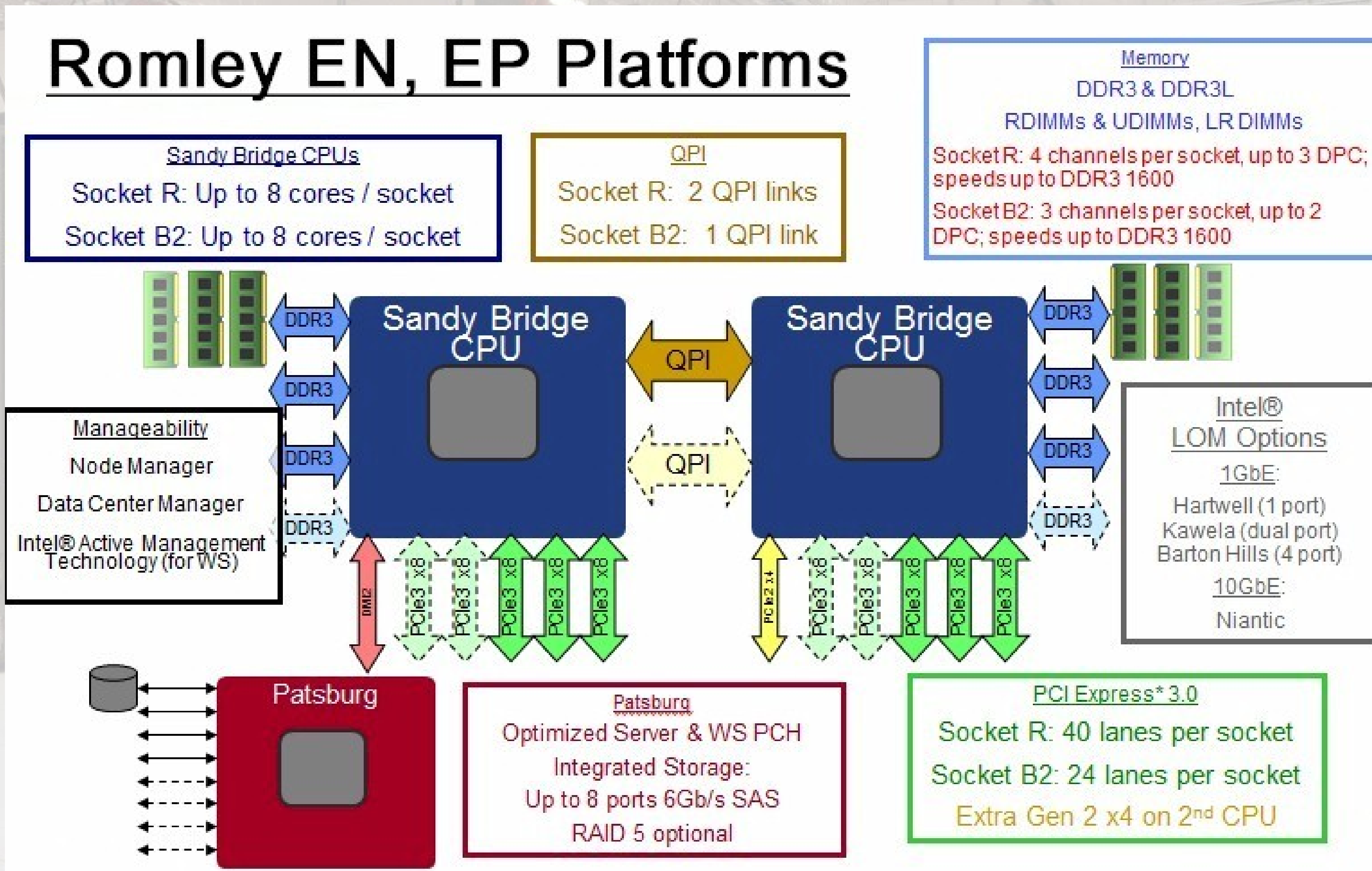
Dynamic processes

Parallel I/O

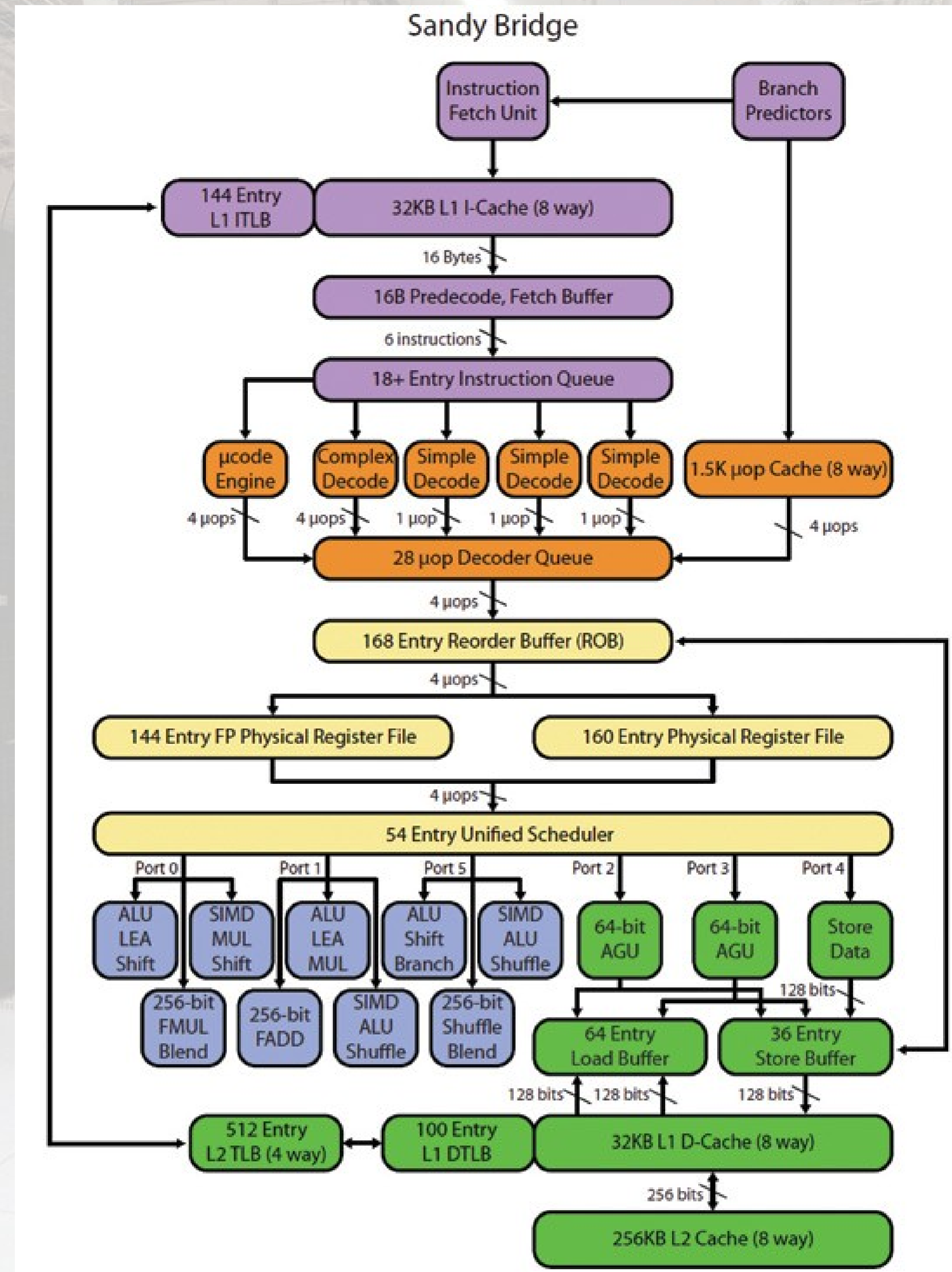
# Intel Tick-Tock

| Intel® Core™ Microarchitecture | | Intel® Microarchitecture Codename Nehalem | | Intel® Microarchitecture Codename Sandy Bridge | | New Intel® Microarchitecture | |
|---|---|---|---|---|---|---|---|
| Merom | Penryn | Nehalem | Westmere | Sandy Bridge | Ivy Bridge | Future | Future |
| 65nm | 45nm | 45nm | 32nm | 32nm | 22nm | 22nm | |
| New Micro-architecture | New Process Technology | New Micro-architecture | New Process Technology | New Micro-architecture | New Process Technology | New Micro-architecture | New Process Technology |
| TOCK | TICK | TOCK | TICK | TOCK | TICK | TOCK | TICK |

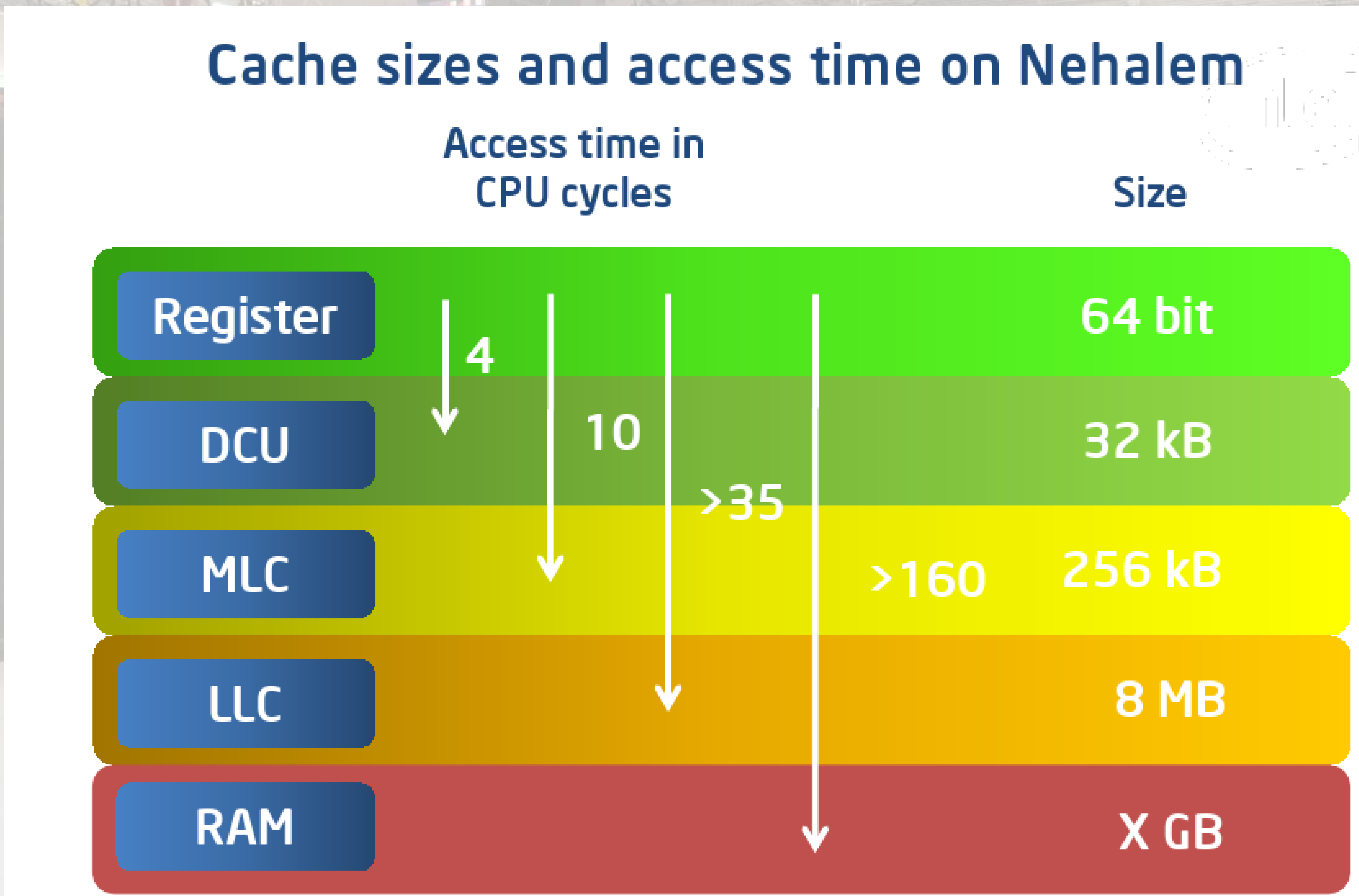
# Intel Sandybridge



# Intel Sandy Bridge Microarchitecture



# Cache Latencies



NSC

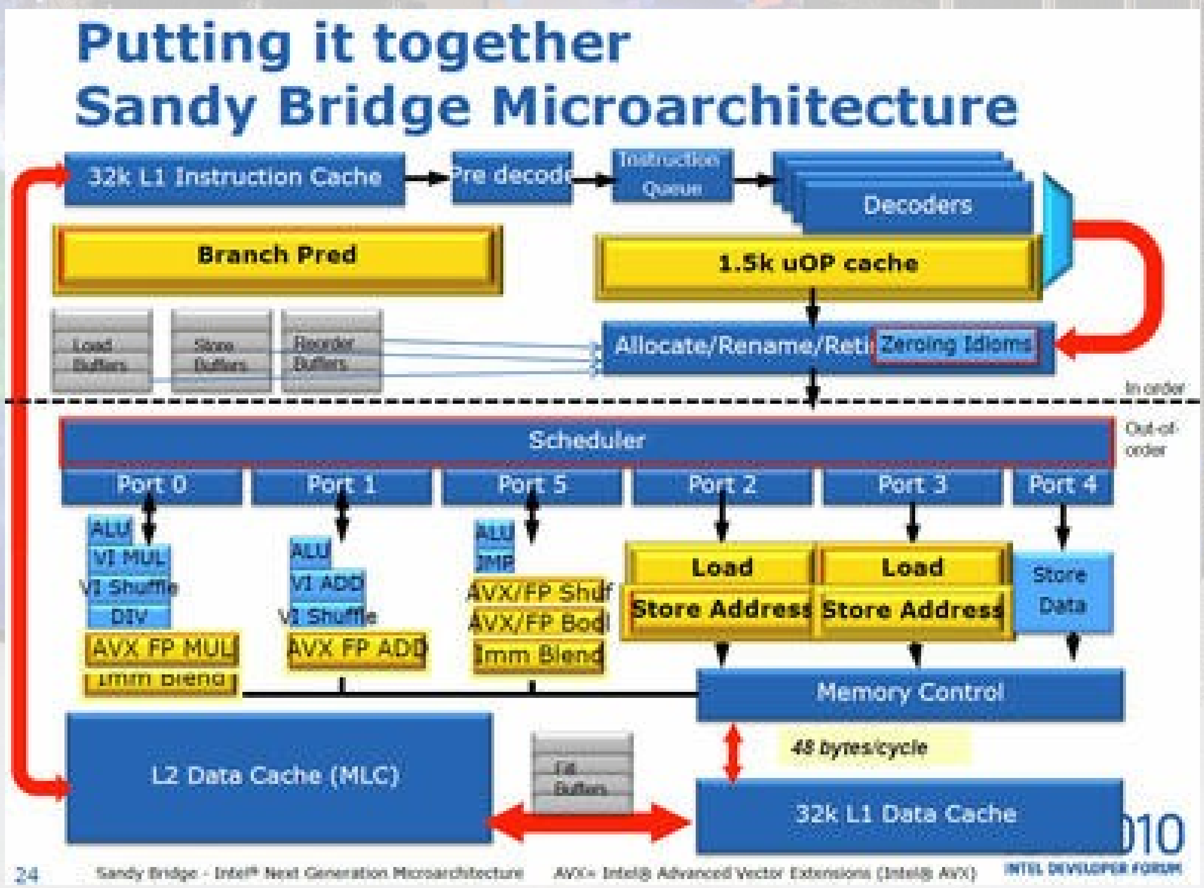
# Intel SandyBridge (Xeon E5)

- Four memory channels (DDR3-1600) on socket R
- AVX – 256 bit vectors
- Larger L3 cache – up to 20 MB
- (2), 4, 8 cores
- Turbo Mode (aggressive over and underclocking)
- TDP: ~ 80-130 W
- 1.8 – 3.6 GHz (Turbo: 4.0)

# Microarchitectures

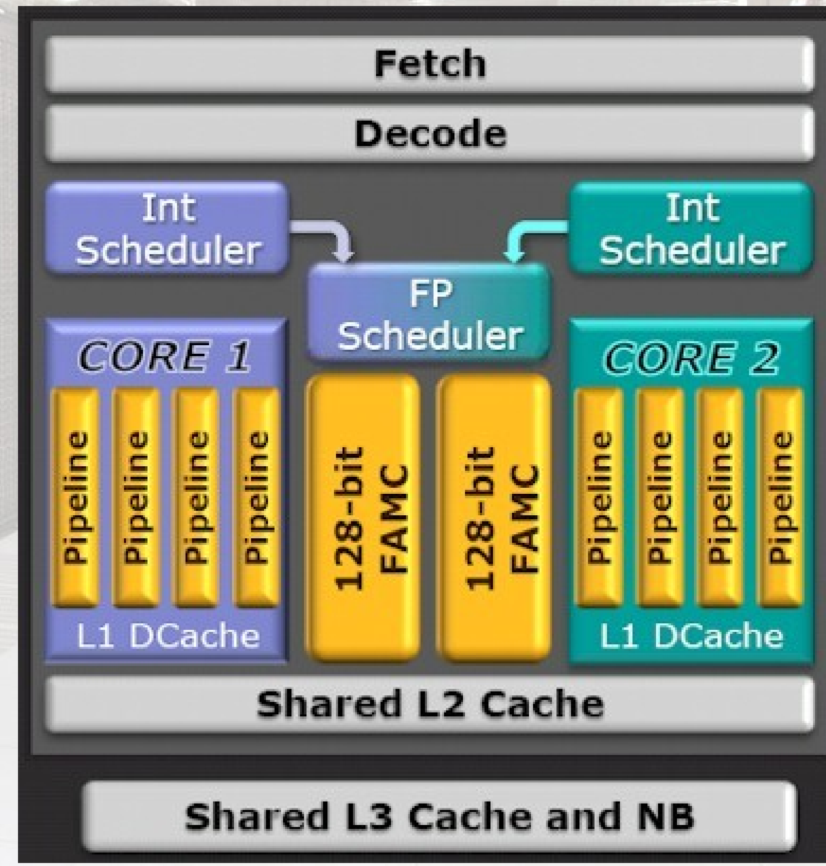
**Intel SandyBridge**

One core




**AMD Bulldozer**

One module, two cores




NSC
National Supercomputer Centre
Network · Storage · Computing

# National Supercomputer Center in Sweden

- Provider of leading edge supercomputing resources to NSC partners SMHI and SAAB and to members of academic institutions troughout Sweden.
- The SNIC*-center at Linköping University
- Independent organisation within Linköping University
- Staff of 30 people
- Created 1989 when Linköping University purchased a Cray XMP in collaboration with SAAB.

*) Swedish National Infrastructure for Computing

# History of NSC

- 1983 Saab bought a Cray 1A (second hand) for simulation of JAS. Research Council bought 2000 h/year for academic research. (1 Mw)
- 1989 NSC started. Cray X-MP/48 (4 proc, 8 Mw)
- 1993 Cray Y-MP/464 (4 proc, later 8, 64 Mw)
- 1995 Cray C90 (6 proc, 256 Mw)
- 1997 Cray T3E (new) (272 proc, 45 GB)
- 1999 SGI Onyx2 at LiU
- 2000 SGI 2K

Clusters

- 1999 Ingvar
  (no 3)
- 2002 Monolith, Bris, Maxwell
  (10, 11, 13)
- 2004 Blixt
  (20)
- 2006 Darkstar
  (25+26)
- 2007 Neolith,Bore/Gimle, Skylord
  (27, 29+31+32, 30)
- 2010 Byvind
  (37)
- 2012 Triolith, Krypton, Skywalker
  (40,39,38)

Underlined = Used for NWP

NSC

# Major Partners and Funding Organisations

**Swedish National Infrastructure for Computing**

**www.snic.vr.se**

Meta-center for six supercomputer centers in Sweden:
**NSC, PDC, HPC2N, UPPMAX, C3SE, LUNARC**

**Swedish Meteorological and Hydrological Institute**

**www.smhi.se**

**Swedish Aeroplane AB**

**www.saabgroup.se**

NeIC

**Nordic e-Infrastructure Collaboration**

**Linköping University**

# Examples of Services

- SNIC General Purpose Computations
- Targeted Research Areas:
  - Electronic Structures
  - Climate
- Numerical Weather Prediction
- WLCG Tier1&2 storage and computation
- Meteorological Archive and Retrieval System – MARS
- Deployment and Optimization of Software/Applications
- SNIC Infrastructure services
- Security Planning and Forensics

# Hardware Resources
## (approximate numbers)

**Computing**

- 34 000 processor cores
- 520 Teraflops (peak)

**Disk Storage**

- 4200 drives (compute servers uncounted)
- 7 Petabyte (raw) ≈
  5.3 Petabyte user space
- GPFS, Lustre, dCache

**Tape Storage**

- 4200 slots
- 3300 tapes (LTO5 and LTO4)
- 13 drives
- 2.5 Petabyte (raw)

**External Network**

- Redundant 10 Gb via LiU to SUNET
- Redundant 1 Gb to SMHI
- 10 Gb to WLCG Nordic Tier-1

# Visit to NSC's computer room

When?

# Triolith – Sweden's fastest computer

# Triolith Hardware

**Compute Servers**

- 1200 HP SL230s Gen8
  - 2 x Intel E5-2660 (2.2 GHz)
  - 1144 “Server A”
    - 32 GiB memory
    - 500 GB SATA disk
  - 56 “Server B”
    - 128 GiB memory
    - 2 x 500 GB SATA disk

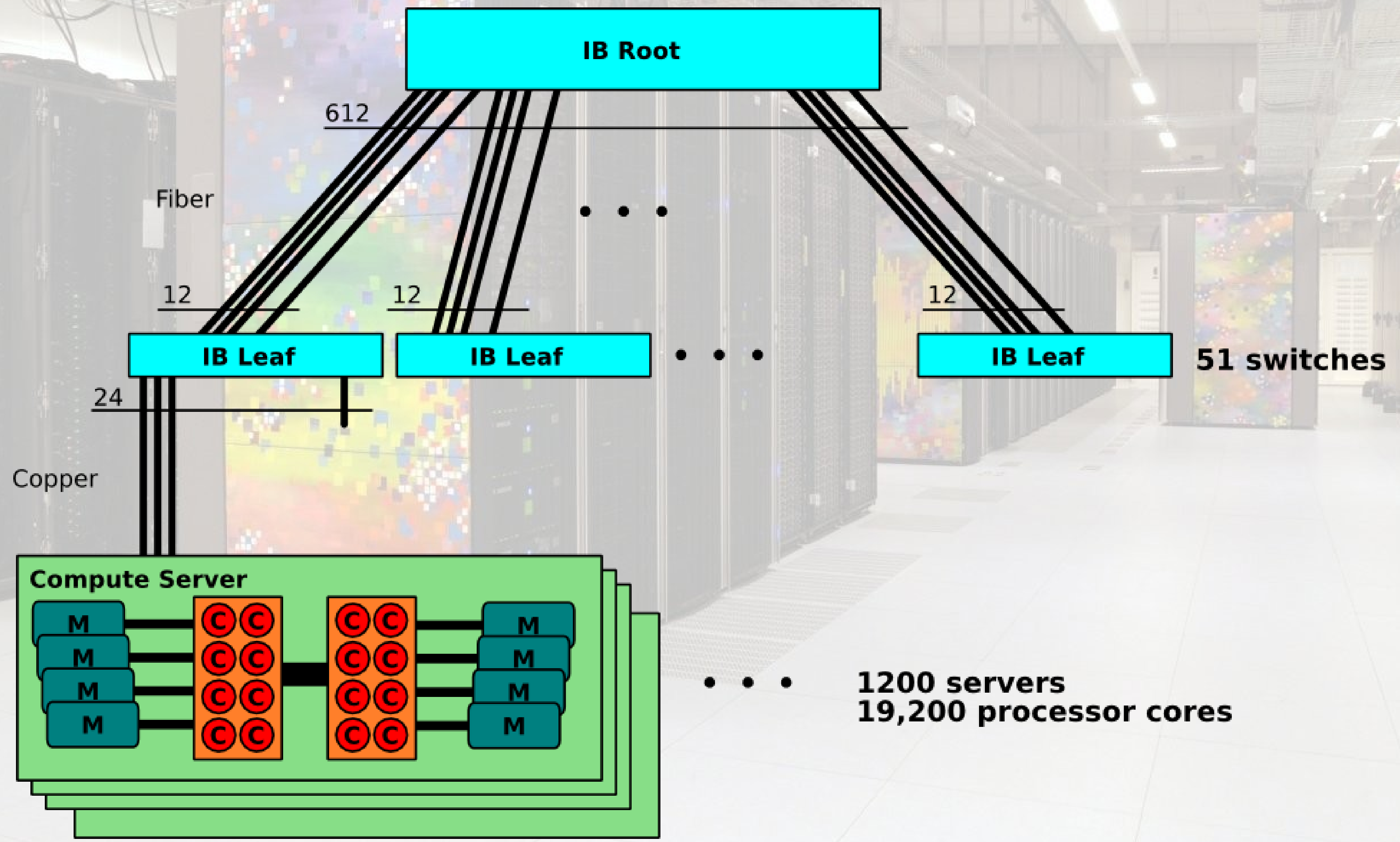
**Networks**

- Mellanox FDR Infiniband
  - One 648p root-switch
  - 51 36p leaf-switches
  - 2:1 blocking
- Gigabit
- iLO

**Auxiliary servers**

- 4 HP DL380 Gen8

# Infiniband Network
# Fat Tree Topology



# Triolith

In total:

- 1 200 HP compute servers
- 19 200 Intel cores
- 150 HP s6500 chassis
- 29 HP racks
- 337.92 Tflop/s (nominal)
- 42.75 TiB memory
- 75 TiB/s aggregate memory BW
- 67.2 Tb network bisection BW

Funded by Swedish Research Council (VR) via SNIC and Linköpings University

Installed and delivered by HP and Go Virtual

# Triolith Linpack Perfomance

Linpack on 1196 servers

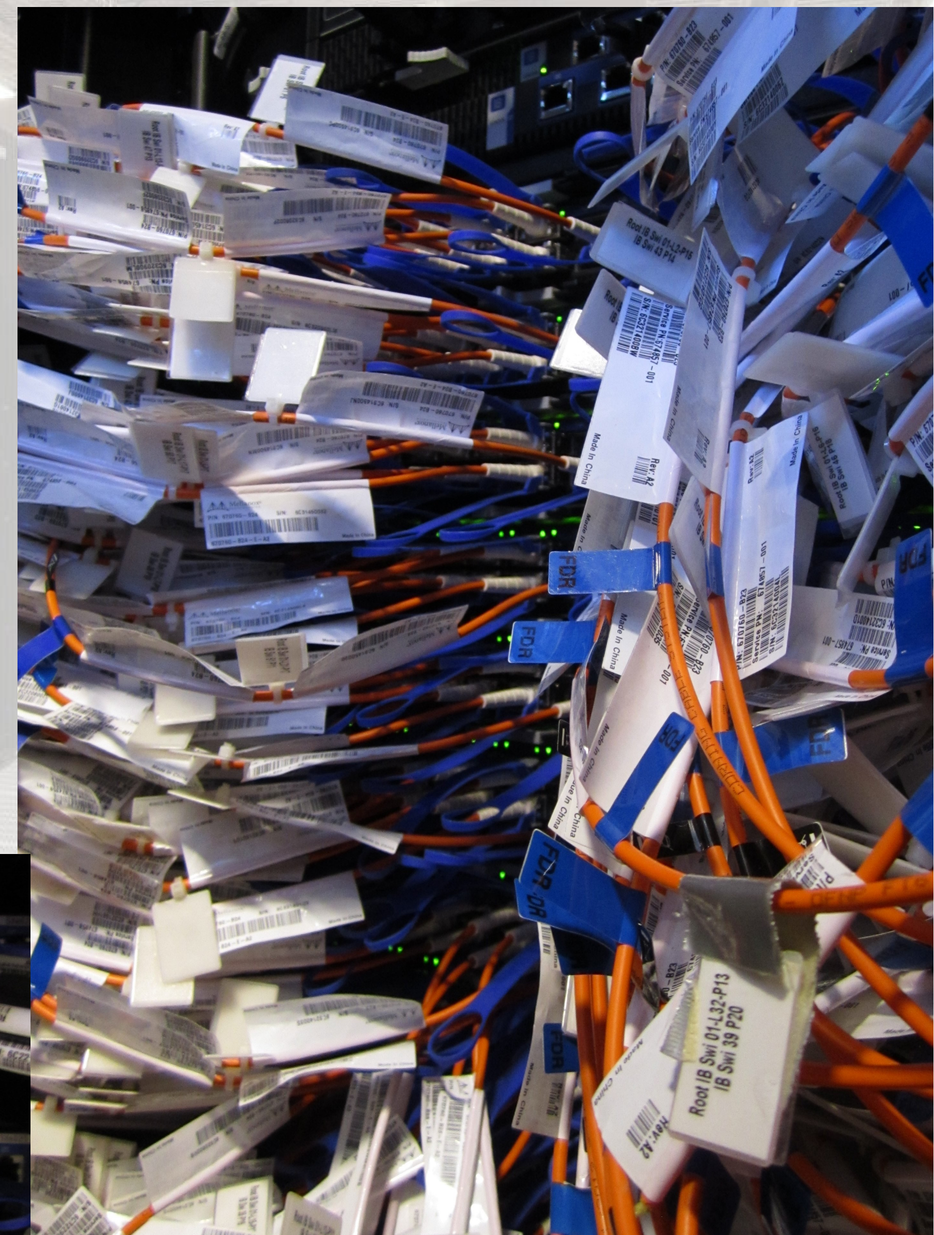
336.8 Tflop/s (nominal)

**303.7** Tflop/s (Linpack)

= 90.2 % efficiency

**#83 on Top500**

380 kW during Linpack = 0.80 GFlop/J



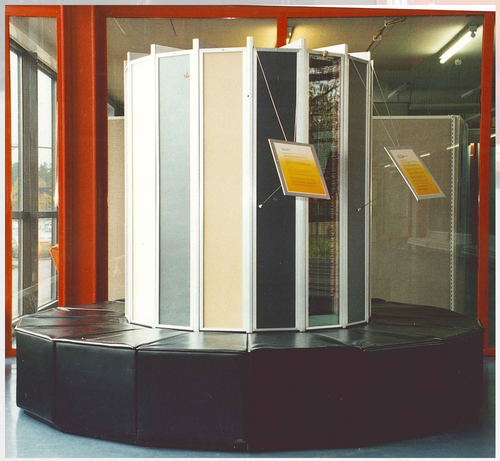

# Performance of one Triolith ≈

**8 Neolith** (2007 – 2012)

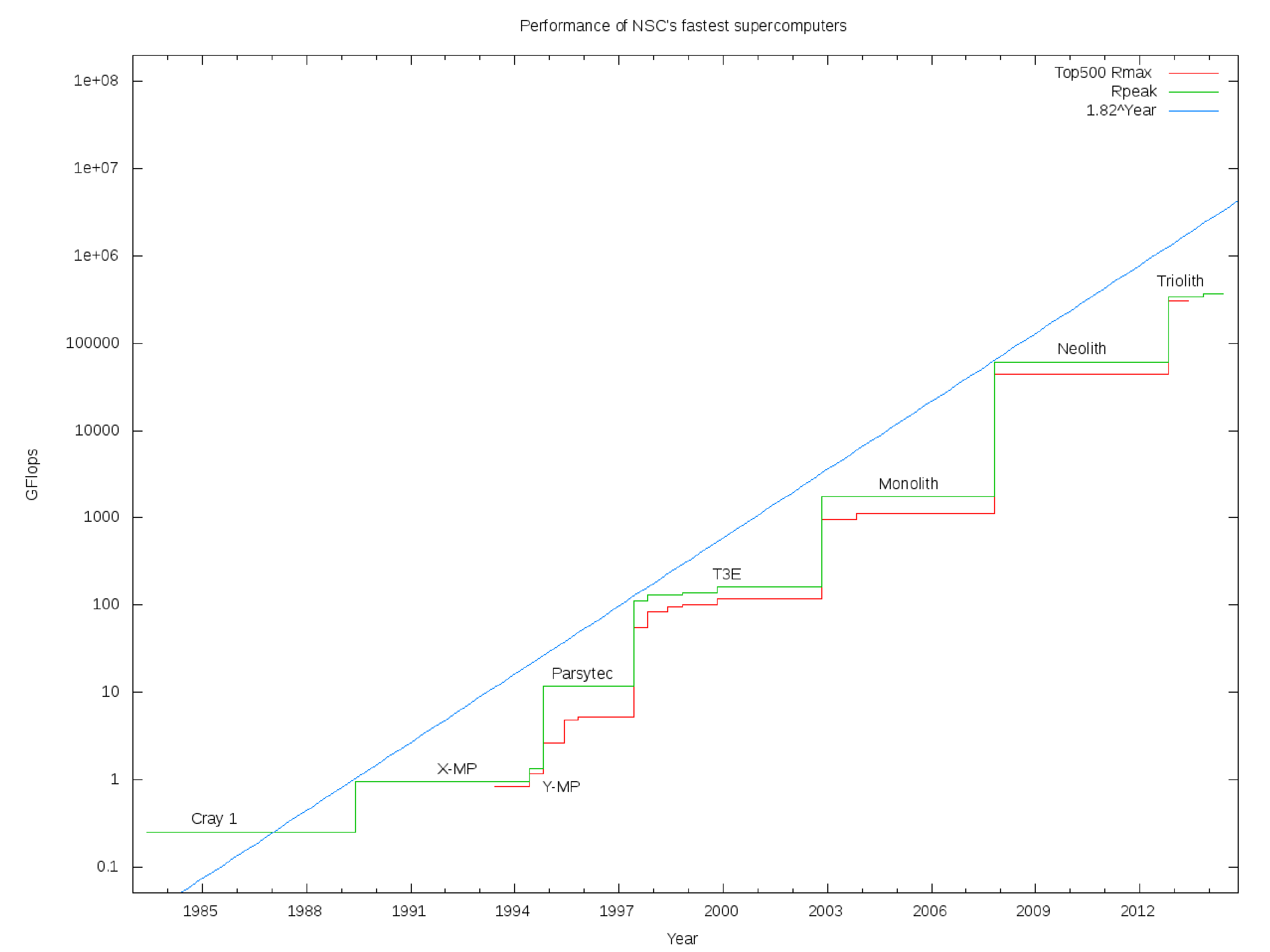
**300 Monolith** (2002 – 2007)

**5,000 Powerful PCs**
(of today)

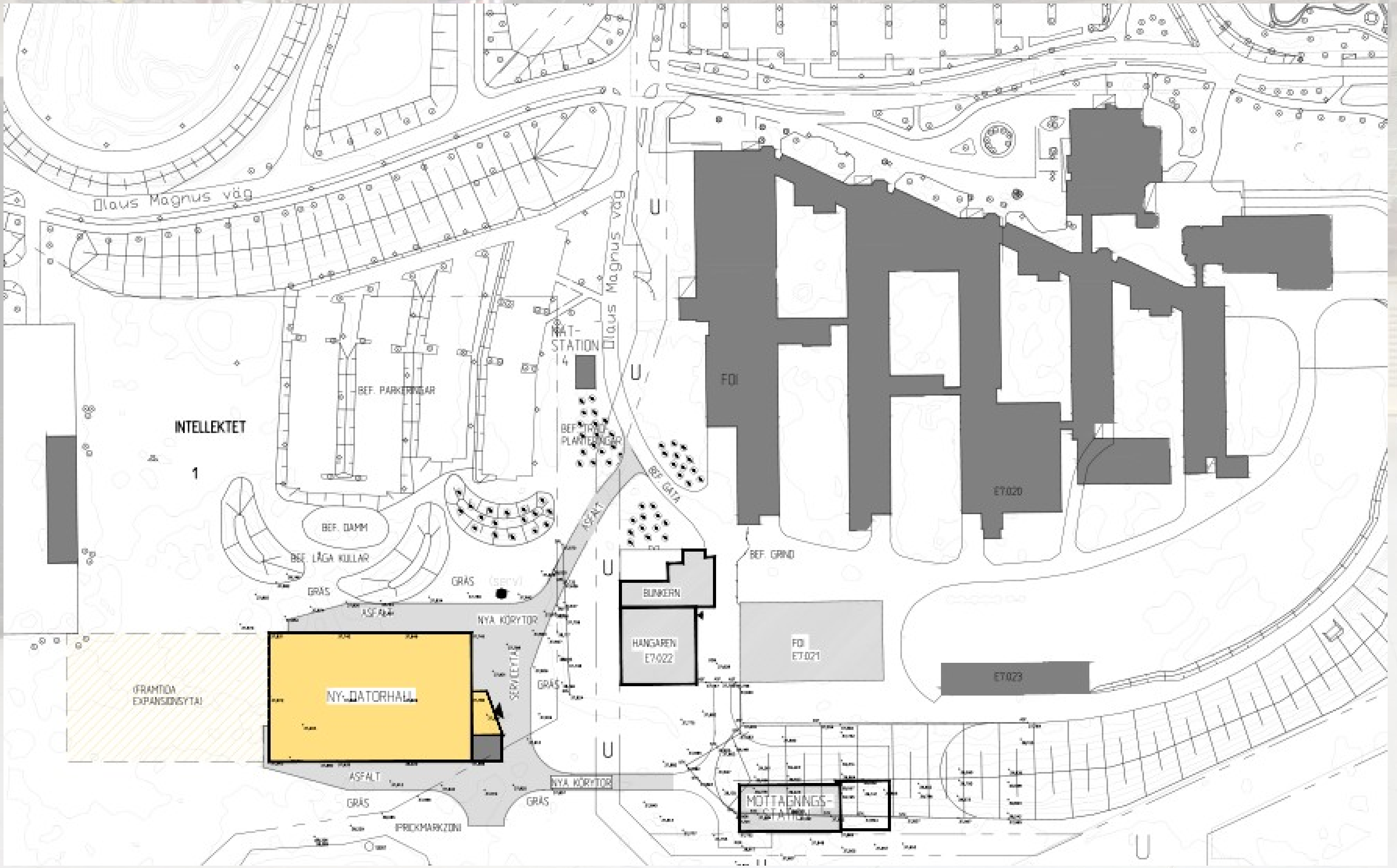
**2,000,000 Cray-1A**
(1983 – 1989)



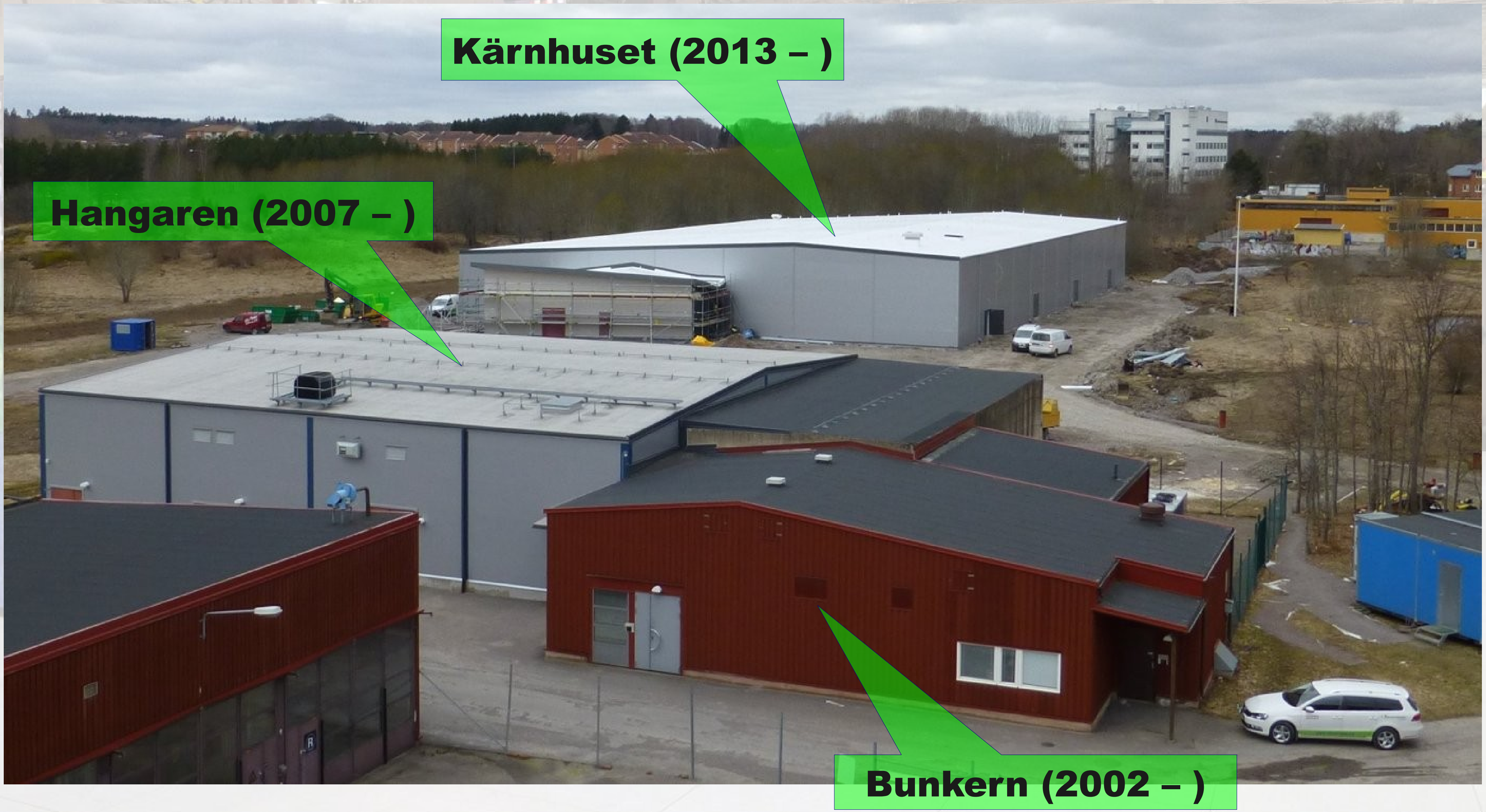
**7,000,000,000 people**
**performing 43,000 flop/s**


Performance of NSC's fastest supercomputers
Top500 Rmax
Rpeak
1.82^Year
GFlops
1e+08
1e+07
1e+06
100000
10000
1000
100
10
1
0.1
Cray 1
X-MP
Y-MP
Parsytec
T3E
Monolith
Neolith
Triolith
1985
1988
1991
1994
1997
2000
2003
2006
2009
2012
Year

# New Computer Room: Kärnhuset



# Computer Room Facilities



# Existing Computer Rooms

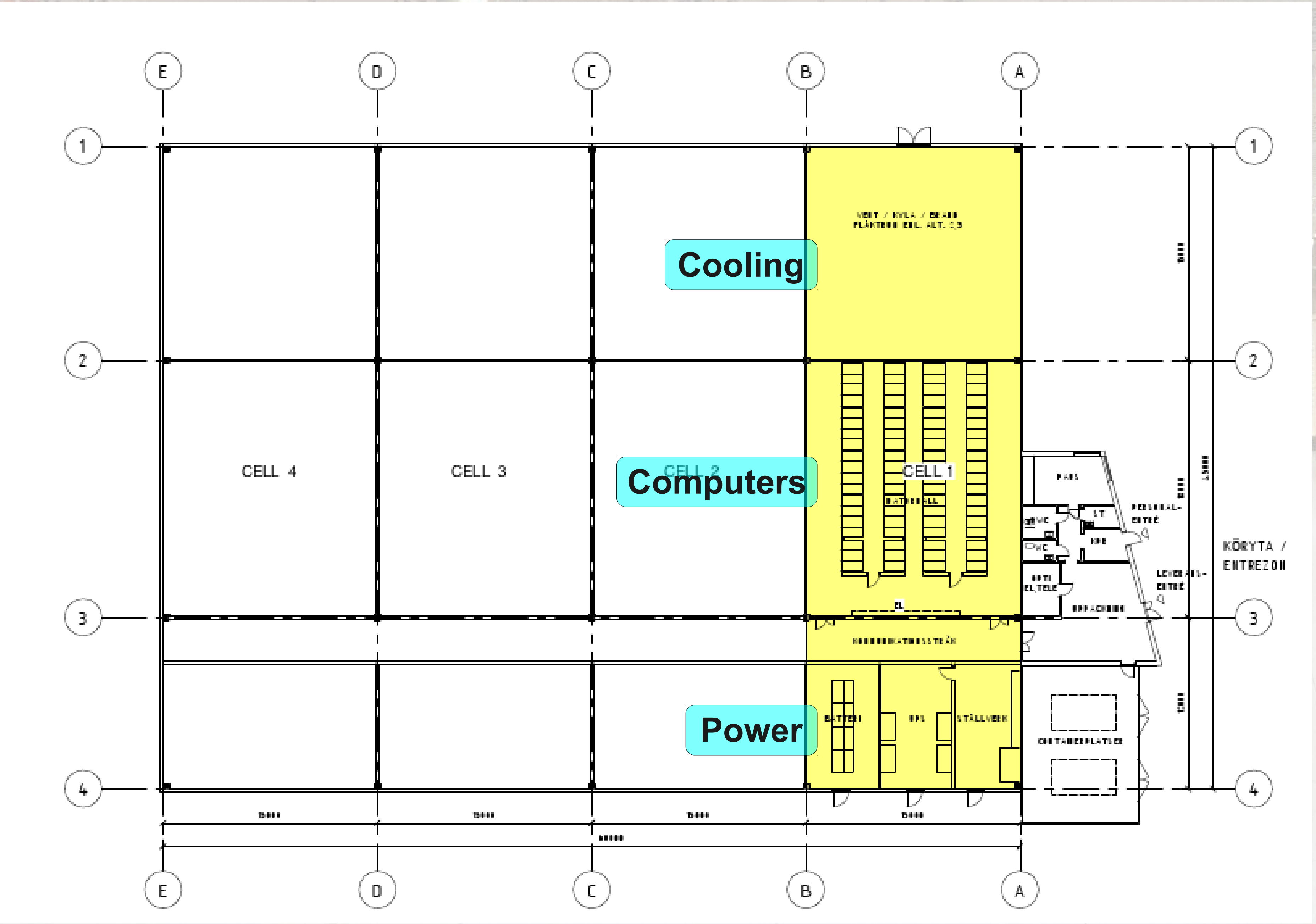
**Bunkern** (2003 – )

- 120 m² floor space for computer equipment
- Installation floor
- Air cooling
- Chiller and air-side economizer
- Fire suppression: water mist
- Max power and cooling: approx: 160 kW
- UPS: 10 minutes
- PUE: 1.28

**Hangaren** (2007 – )

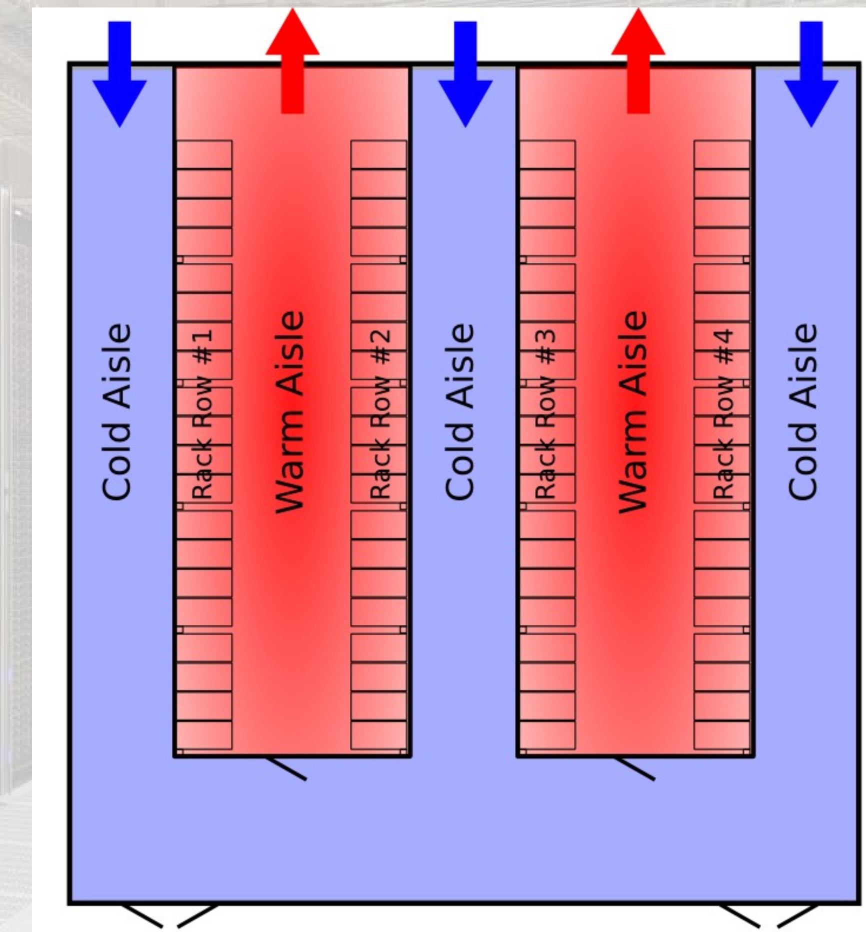
- 240 m² floor space for computer equipment
- Installation floor
- Both open air cooled systems and water cooled encapsulated racks.
- District cooling
- Max power and cooling: approx. 840 kW
- UPS: 100% for 4 min + 15 % for 1 h
- PUE: 1.17
  - Power dist., UPS: ~ 8%
  - Pumps and Fans: ~ 6%
  - Auxiliary: ~3%
  - No power for producing cold water

# New Computer Room Building (2013 – )



# Kärnhuset, Cell #1

- Max 1 MW computer load
- Max 80 racks
- Air cooling
- Aisle separation from the start
- No installation floor
- District cooling
- Ready Summer 2013




Computer room: 280 m²
Air handling: 100 m²
Pipes and pumps: 100 m²
Power distribution, UPS and batteries: 150 m²
Access and transport: 40 m²
Auxiliary: 60 m²

# Kärnhuset, cell #2, #3, #4

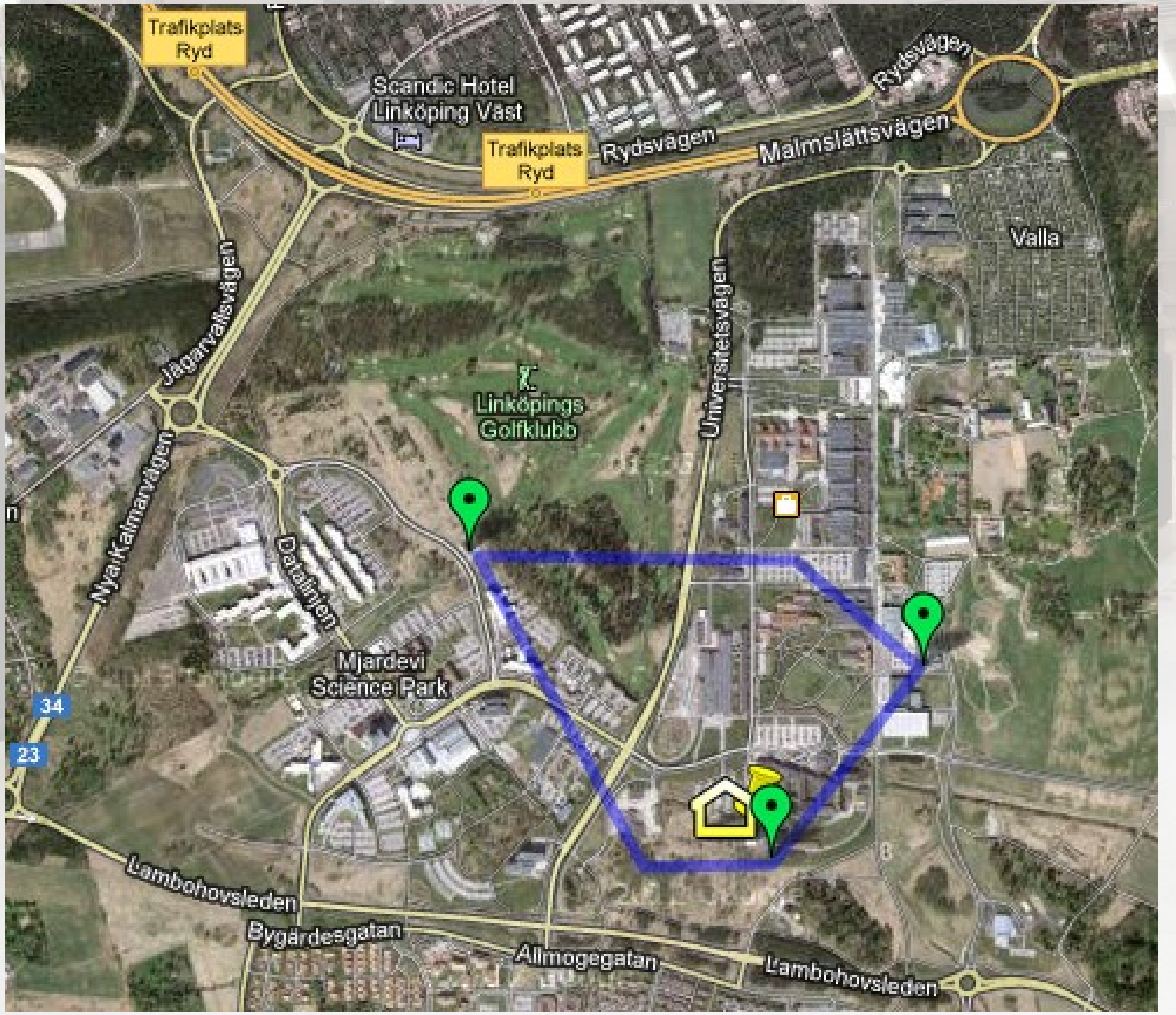
- Water based cooling?
- Container shipping unit?
- …

We are prepared to host more HPC in an efficient environment.

# Distribution Central

- Power distribution, 11 kV switchgear
- District Cooling production, powered by:
  - Summer: District Heating
  - Winter: Cooling Towers
- Run in parallel with two other distribution centrals within the district.

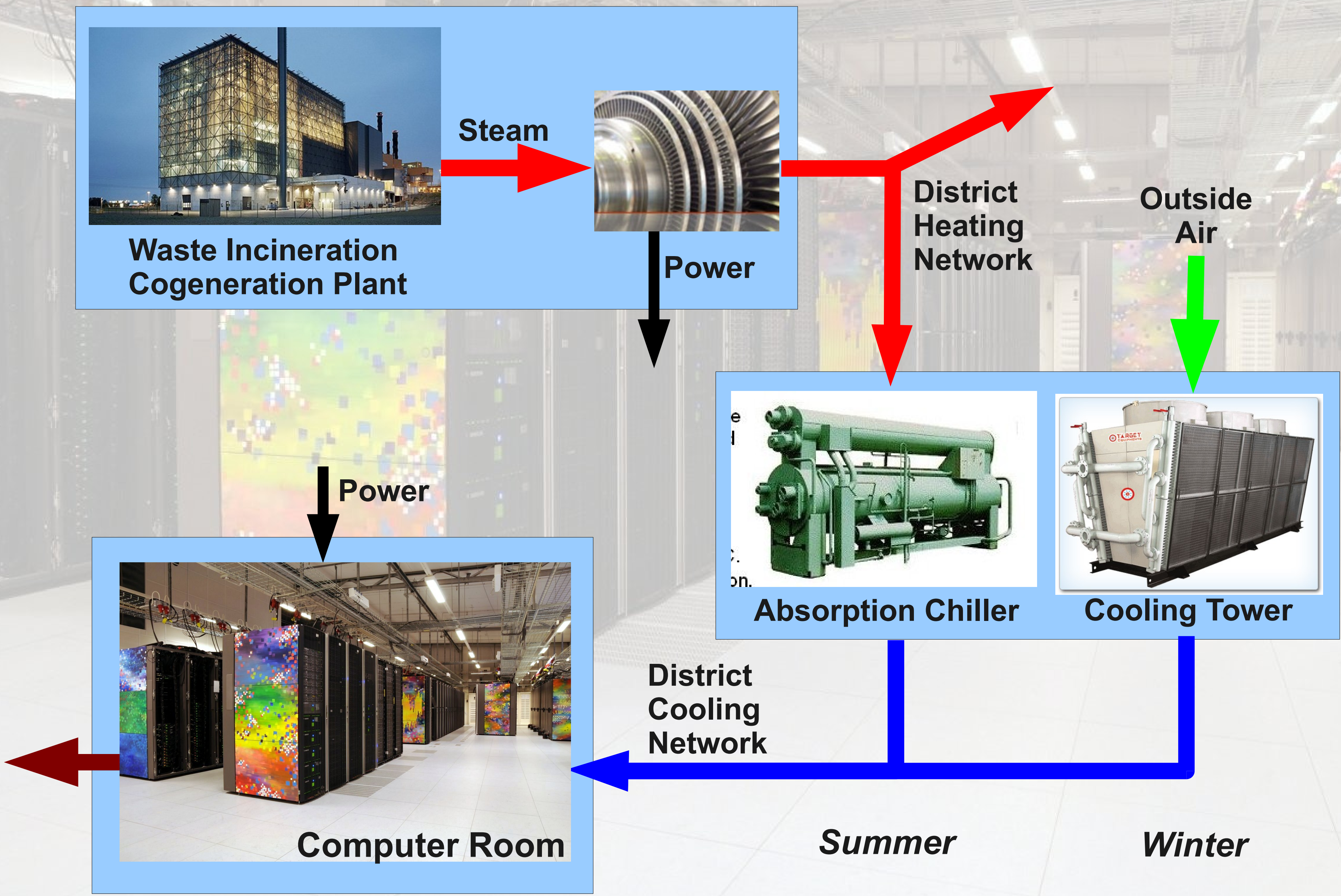
# District Cooling



# Co-generation of power and district heating

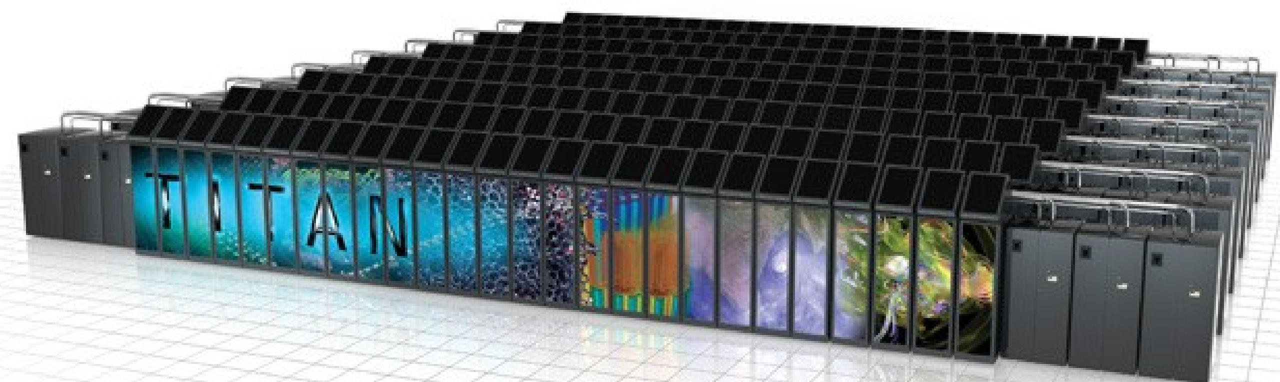

- Waste incineration plan
- In addition: boilers powered with:
  - Biomass
  - Coal & Rubber (energy recycling)
  - Petroleum (peak)
- 38 Hydroelectric plants

# Combined Cooling, Heat, and Power (CCHP)



# No. 1 on Top500: Titan

- 200 cabinets (404 m²)
- 18,688 nodes
  - AMD Opteron 6274
  - 32 GB memory
  - NVIDIA Tesla K20X
- 299,008 cores (CPU)
- 700 TiB memory
- 3D Gemini network
- Peak: 27 petaFLOPS
- Linpack: 17,59 petaFLOPS
- 8,2 MW (peak)




# Future: exaFLOPS

Targets by DARPA:

- 2018: 1 exaFLOPS
    - 2008: 1 petaFLOPS, LANL, IBM
    - 1998: 1 teraFLOPS, ASCI Red, Sandia
- 32-64 PiB memory
- ~20 MW
- MTTI: O(1 day)

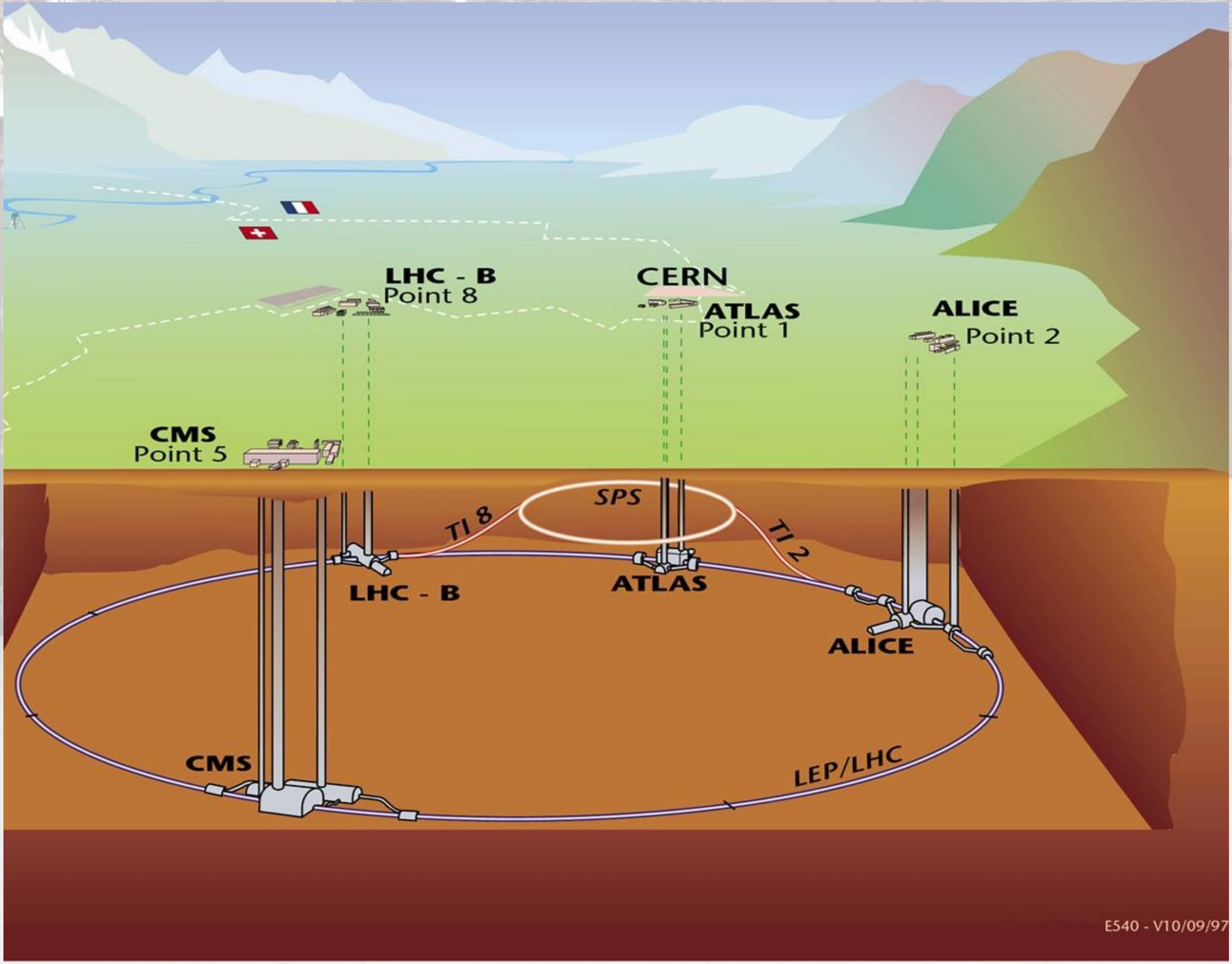
# Applications

# Applications

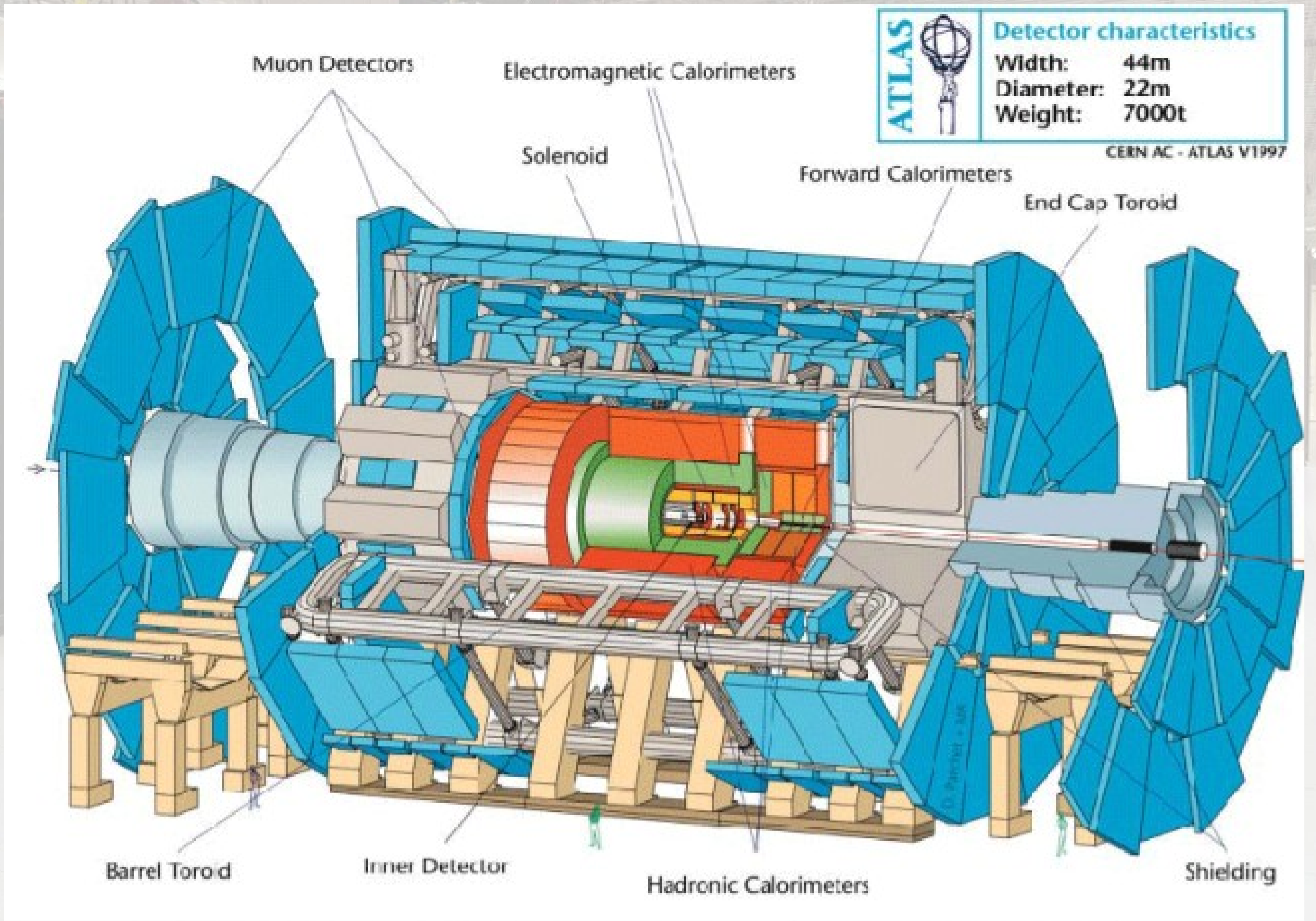
- Climate
  - Extreme weather
  - Carbon, Methane, and Nitrogen cycles
  - $CO_2$ sequestration
  - Scenario replications, ensembles
  - Increase time scale
- Computational Fluid Dynamics
  - Design of aircrafts, vehicles, submarines
  - Combustion, Turbulence
- Advanced materials
  - Solar cells
  - Fuel cells
  - Battery technology
  - Long term storage of Nuclear material
- Bioinformatics
  - Human genome
  - Drug design
- Astronomy
- Nuclear fusion
- Basic Research

# Large Hadron Collider (LHC)

# LHC Experiments



# LHC Experiment: ATLAS

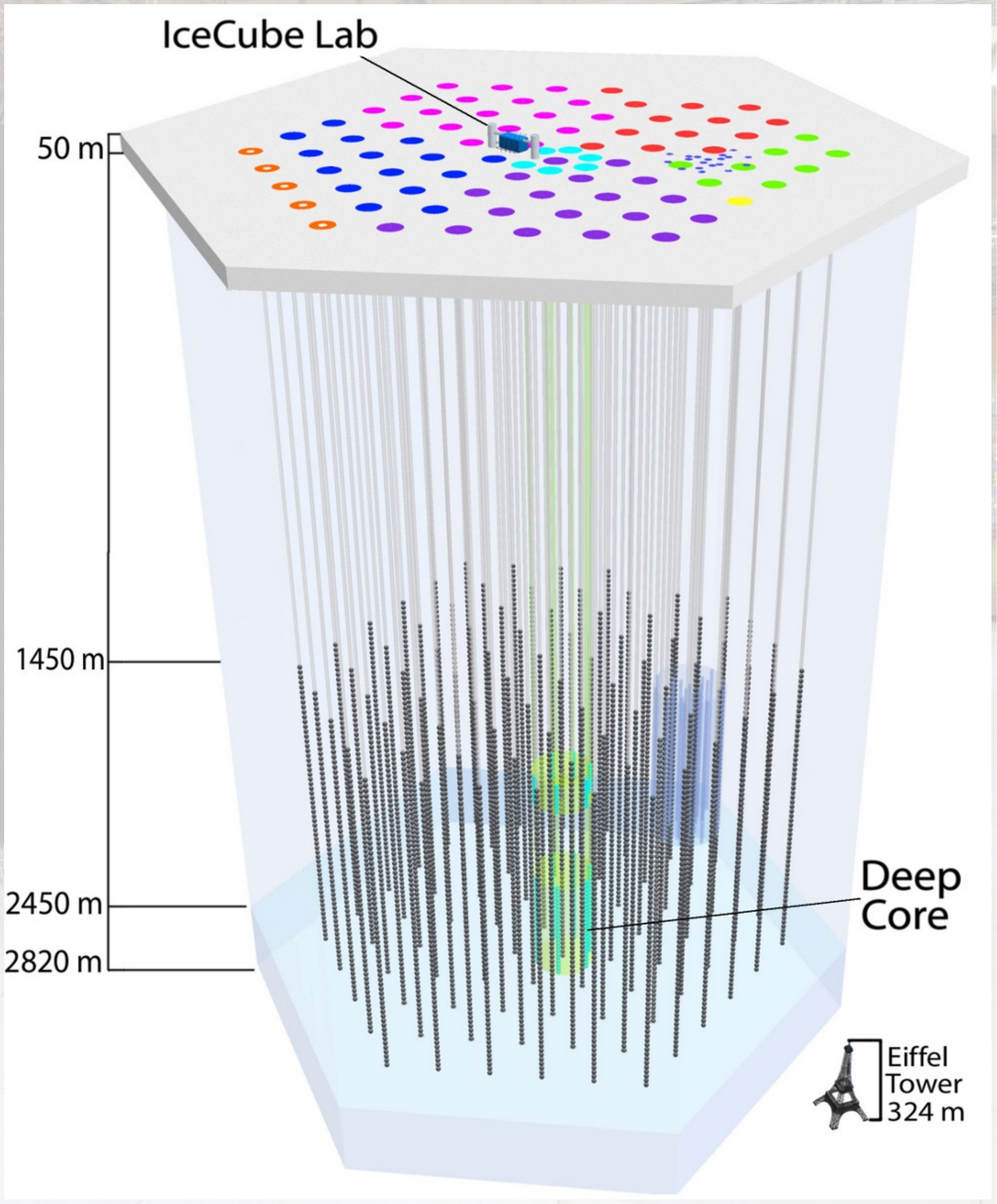


# ATLAS Detector

# IceCube - Neutrino observatoriet

**Klas Hultqvist, Stockholms universitet**

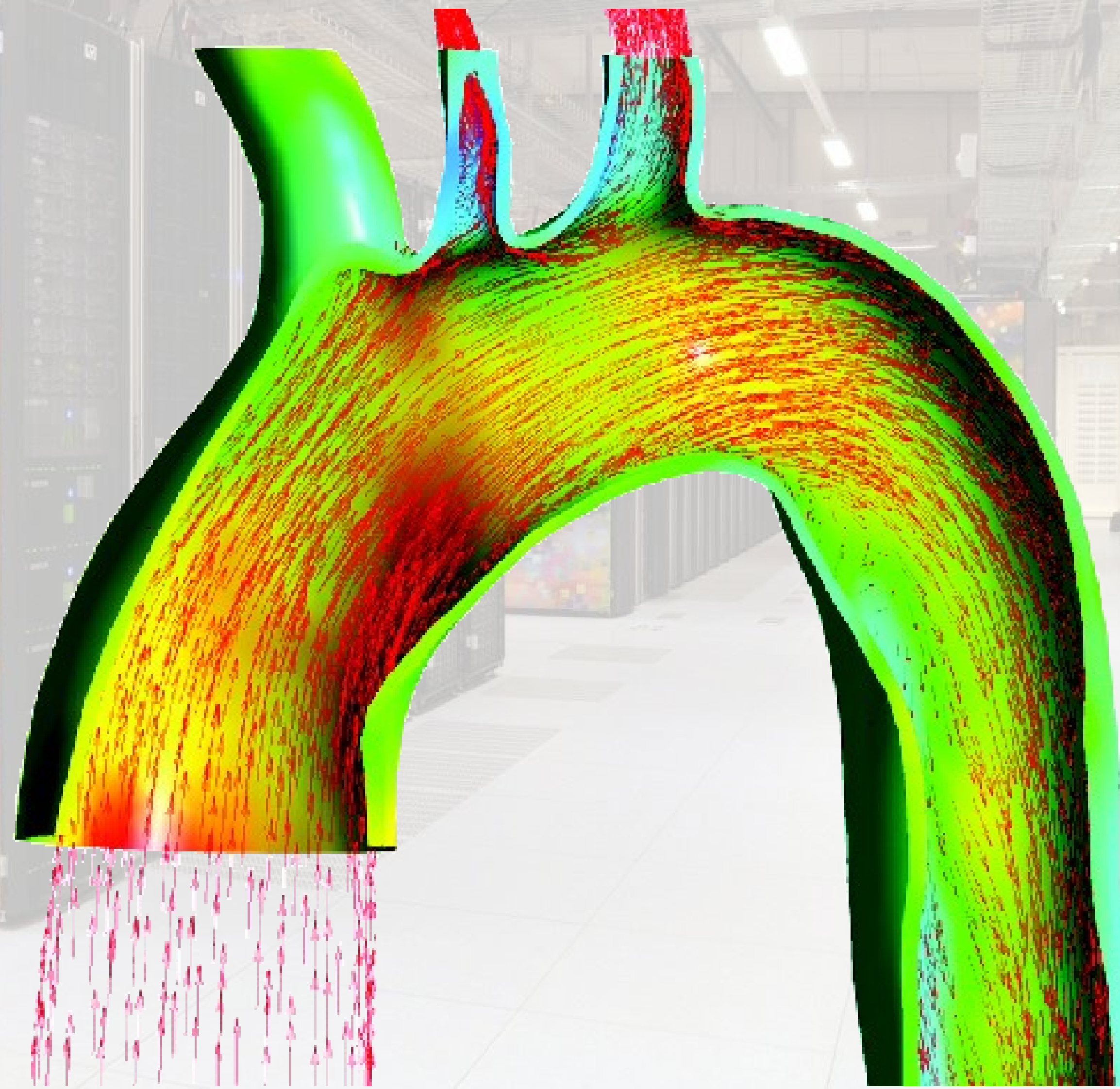
- Detektor vid sydpolen
- 86 vertikala band
- 5160 optiska moduler i 1 km$^3$ is
- Detekterar Čerenkov strålning från sekundära partiklar neutrinos → muons
- 2000 händelser eller 10 MByte data per sekund
- 100 TByte data per år



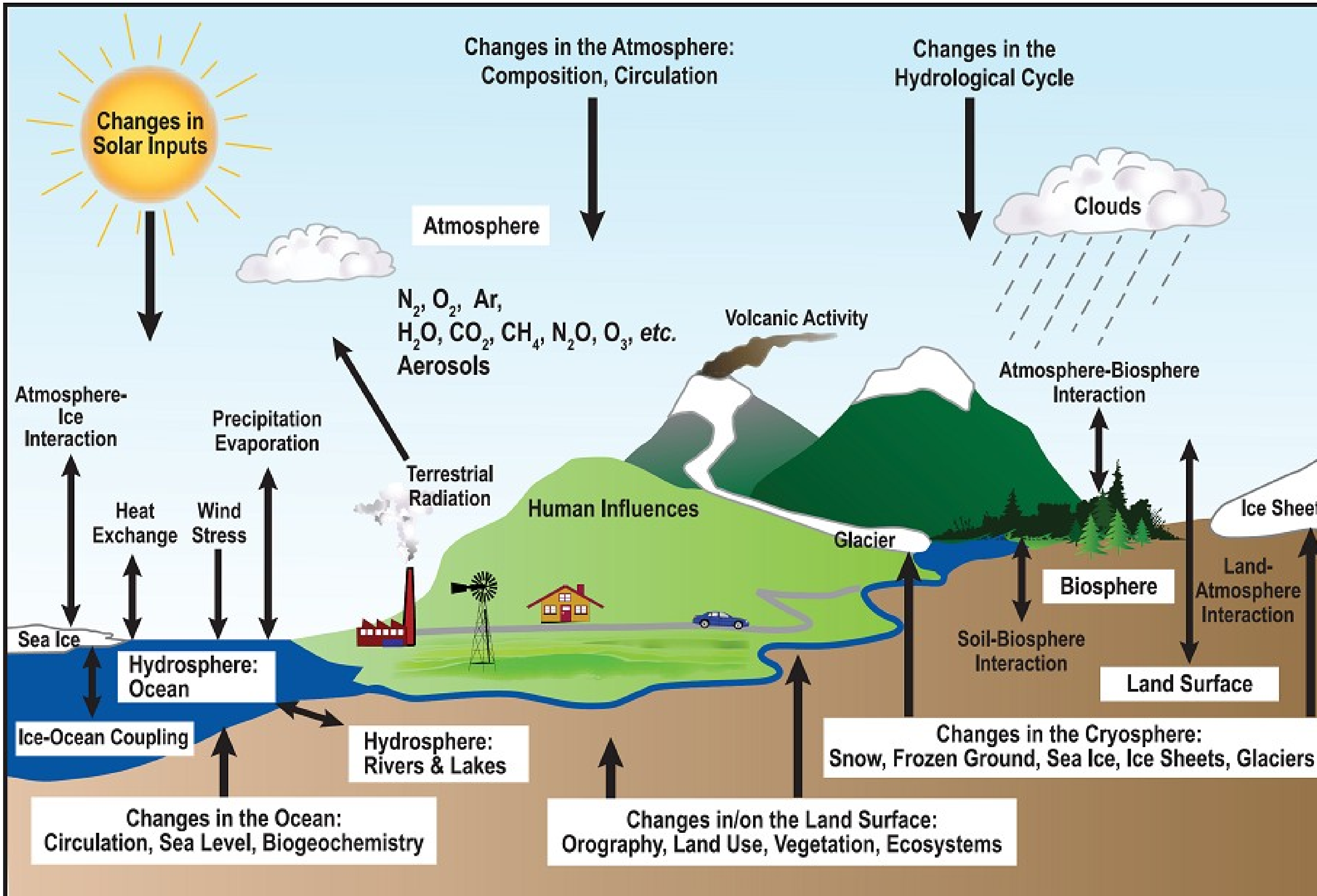
# Simulering av blodflöde i aortan

**Matts Karlsson, Linköpings universitet**

Blodflöde i en mänsklig aorta. Skjuvspänningen i kärlväggarna är färgkodad.



# Climate Simulation

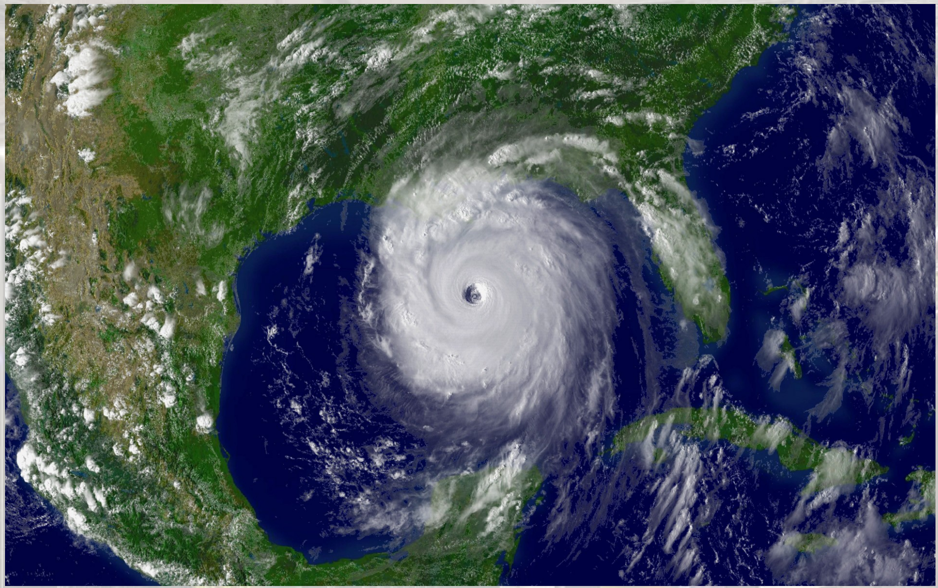



**FAQ 1.2, Figure 1.** *Schematic view of the components of the climate system, their processes and interactions.*

# Numerical Weather Prediction

Challenges

- Stochastic process
- Chaotic nature of fluid dynamic equations
- Predict extreme weather conditions
- Increase in precision and accuracy
- Deadlines

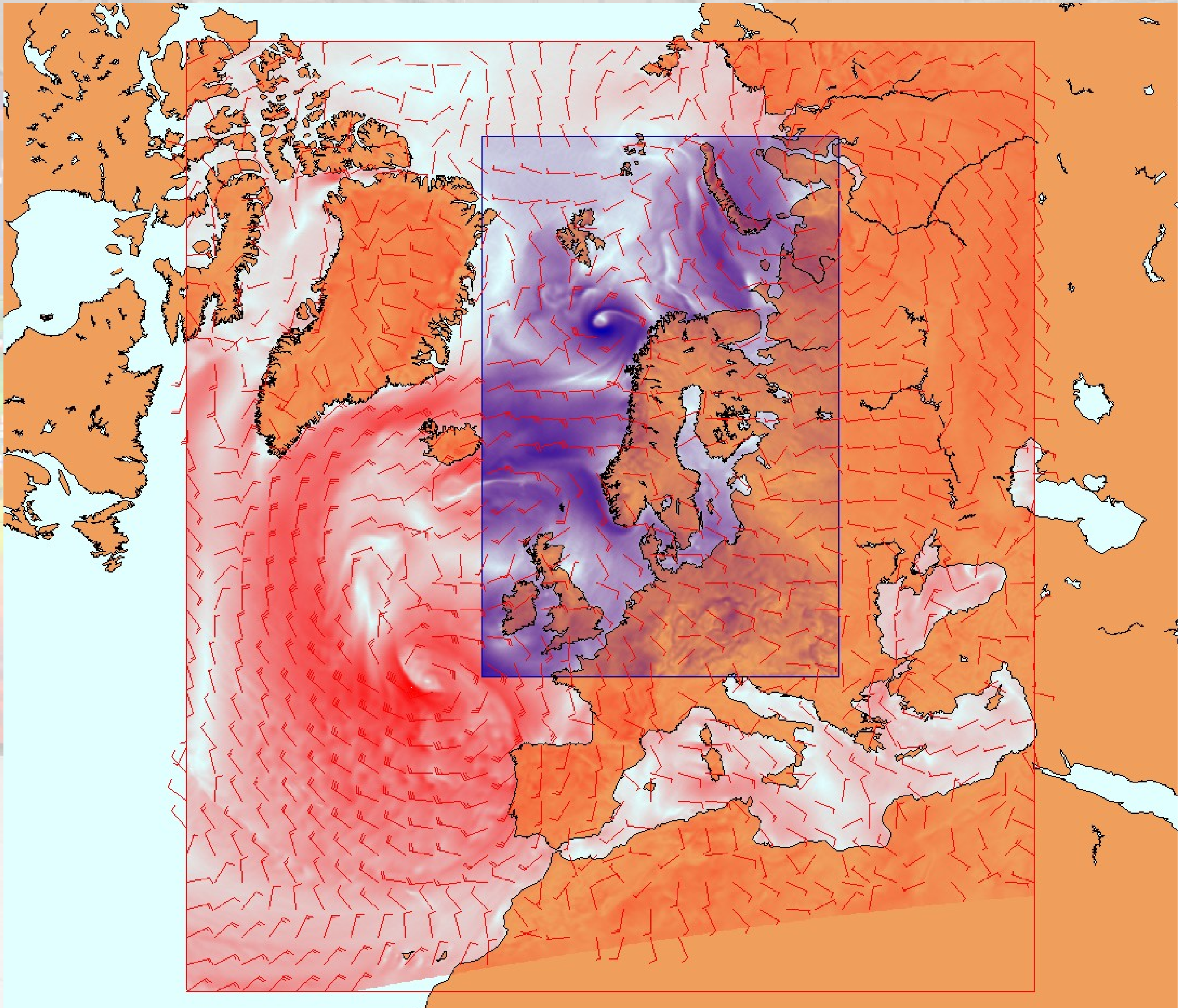


**Hurricane Katarina, 2005**

**Gudrun (Erwin), 2005, Byholma Timber storage**

# Regional NWP (SMHI & Metno)

**1212 x 1360 @ 5,5 km**
**1134 x 1720 @ 2,5 km**
**60-100 levels**




Mellanox
NSC