TSEA28, Datorteknik Y
Guest Lecture
25 April 2012

NSC
National Supercomputer Centre
Network · Storage · Computing

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

Contents

- High Performance Computing
- Processors of today, example: Intel Xeon
- National Supercomputer Centre
- Large scale computing resources
- Applications

Slides: http://www.nsc.liu.se/~nican/education/tsea28_2012.html

What is a Supercomputer?

Cray-1A

What are the differences?


#2 on top500: Tianhe-1A

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, ...

Examples

Matter system at NSC
Nehalem (Intel Xeon 5500)

- 2.26 GHz clock
- 4 Flop / clock / core
- 4 core / socket
- 2 socket / server
- 516 compute servers
- 2.26 * 4 * 4 * 2 * 516 = 37 Tflop/s

Triolith (not installed yet)
SandyBridge (Intel Xeon E5)

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

Hybrid computing

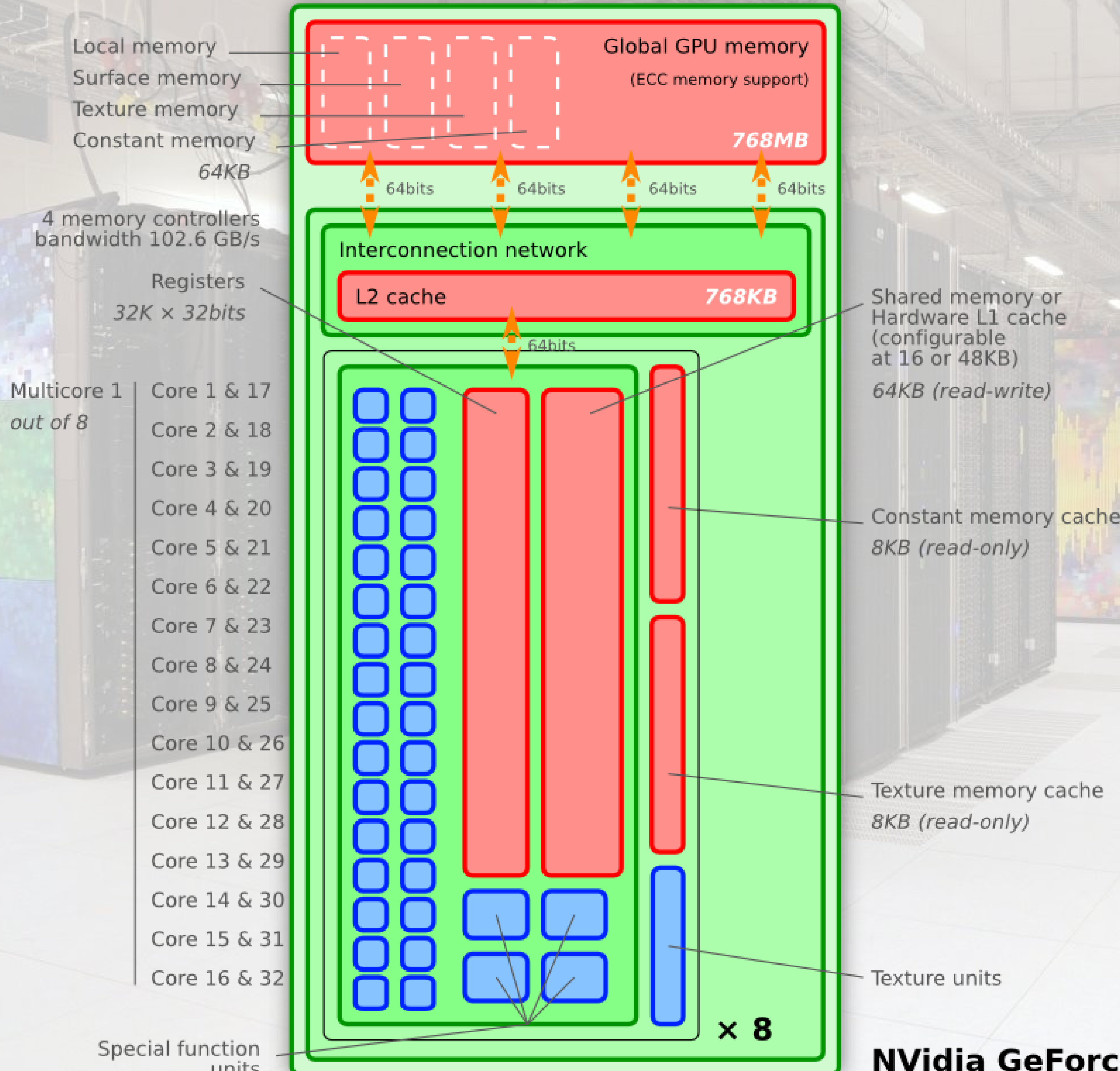
Merge traditional CPU with high performance co-processor

- Today: General-purpose computing on graphics processing units (GPGPU)
- Open Computing Language (OpenCL)
- NVIDIA CUDA

Hardware

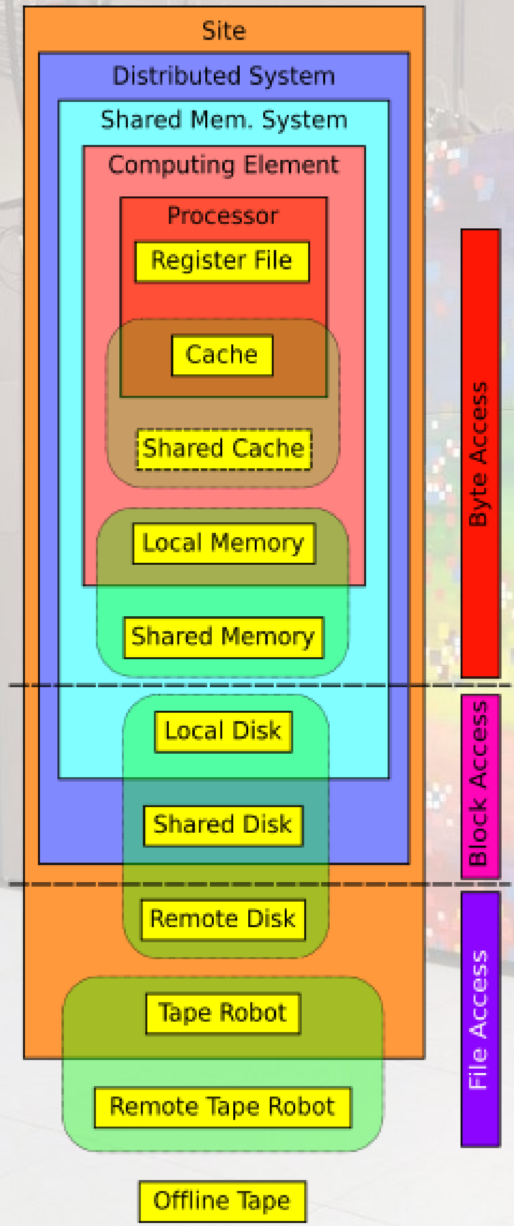
- NVIDIA Fermi, (Kepler)
- AMD Fusion, ATI
- Intel Knights Ferry, Knights Corner, Knights Bridge

NVIDIA Fermi



NVidia GeForce GTX 465

Storage hierachy



- 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

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

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

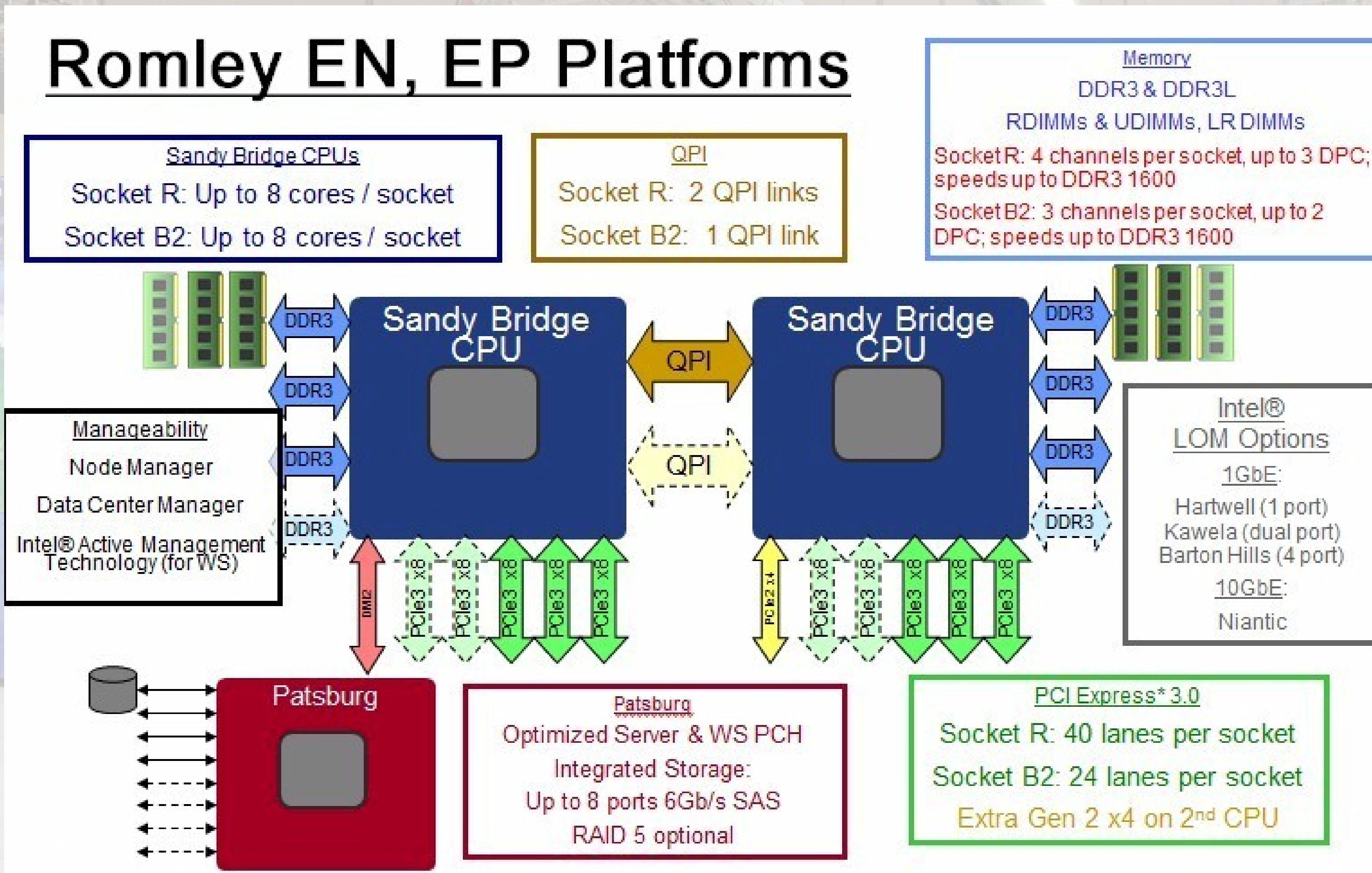
Dynamic processes

Parallel I/O

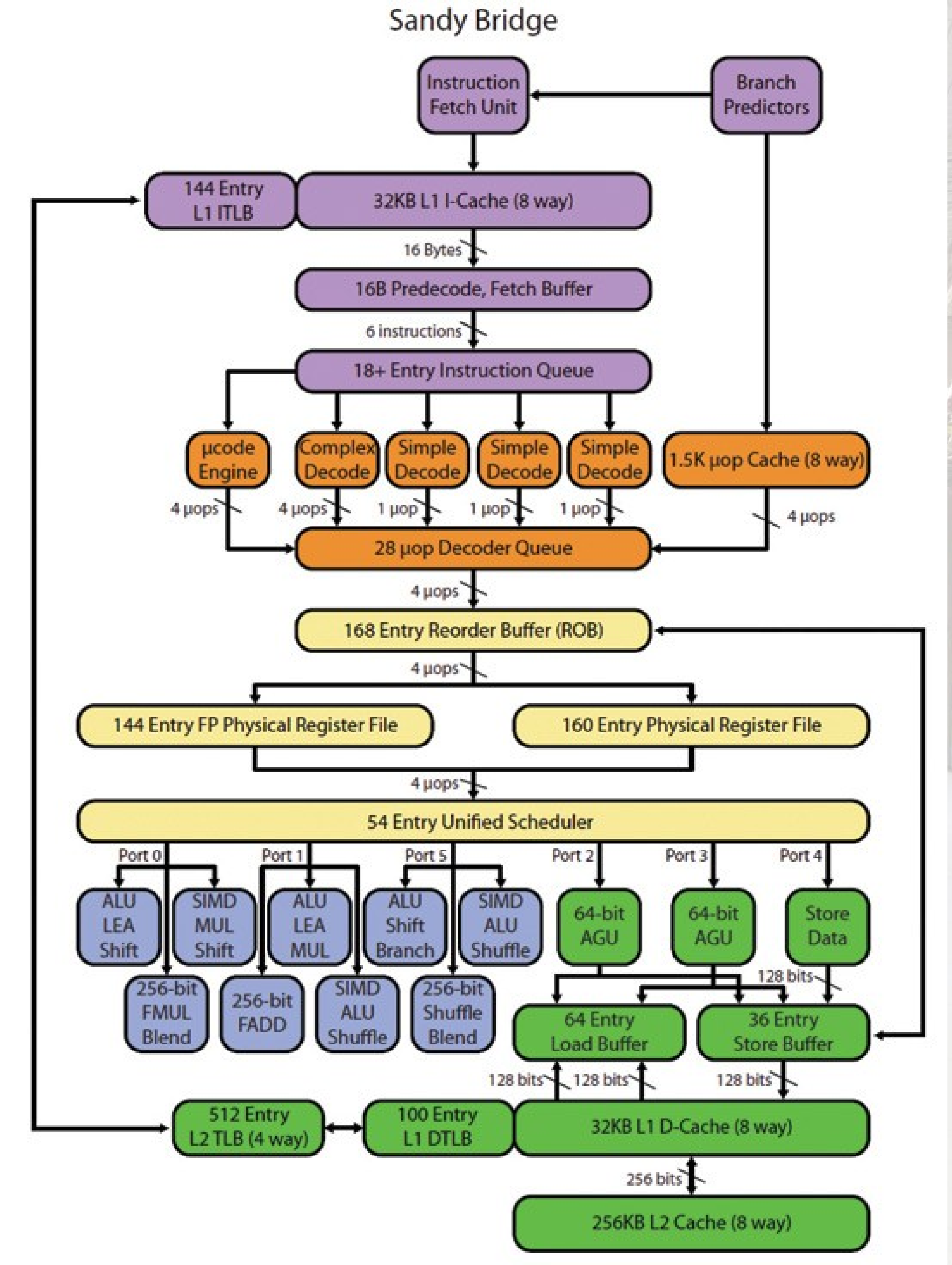
Intel Tick-Tock

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

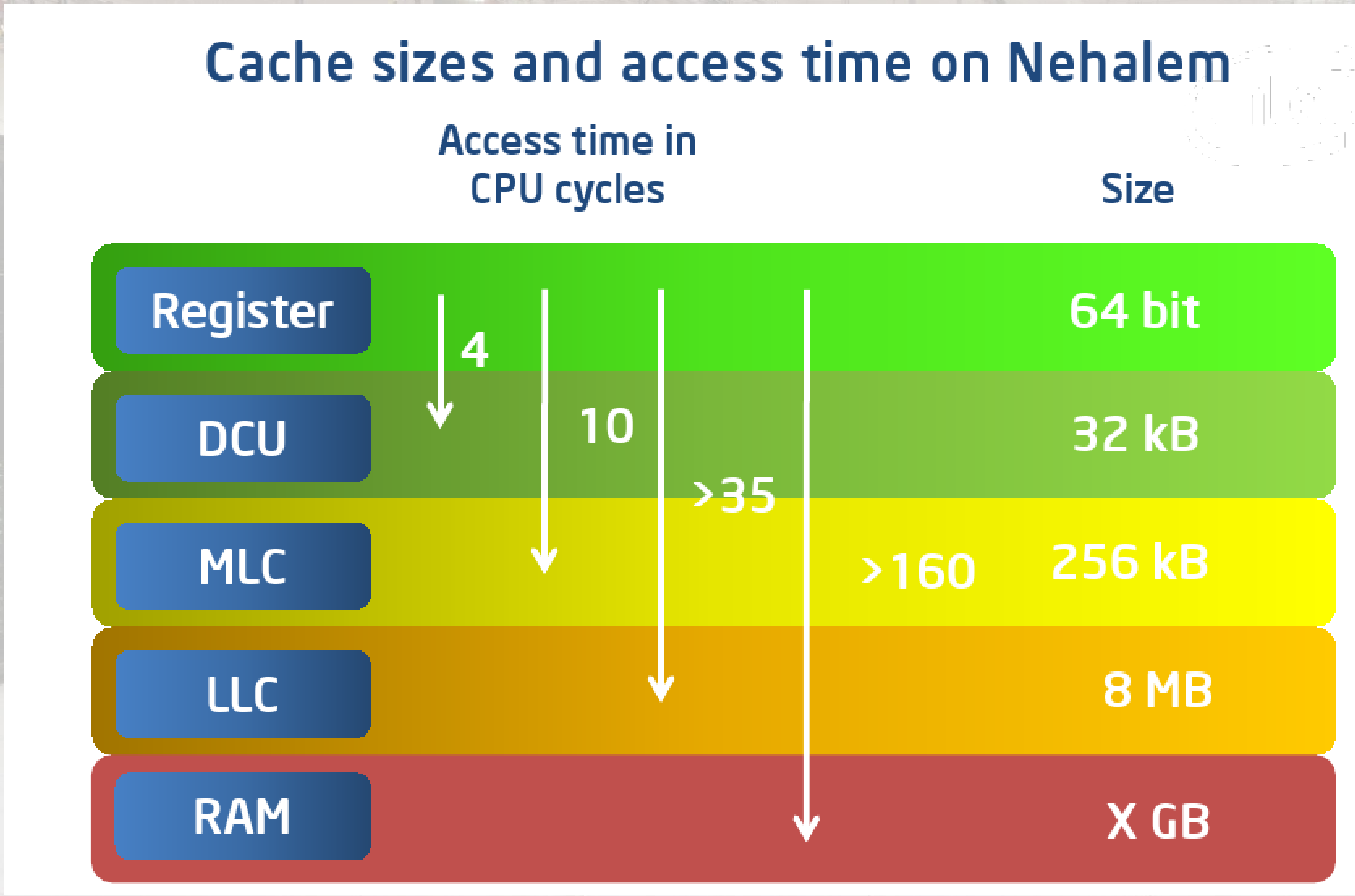
Intel Sandybridge



Intel Sandy Bridge Microarchitecture



Cache Latencies



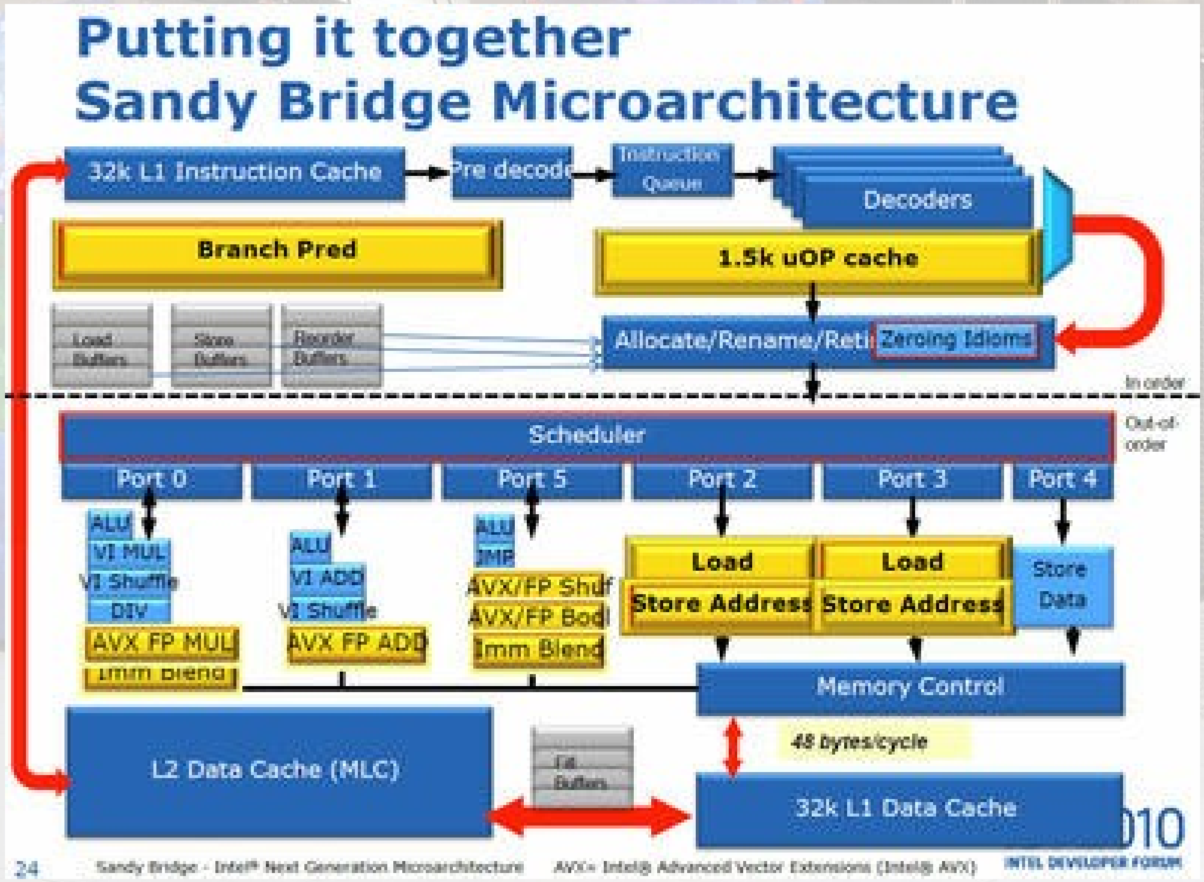
Intel SandyBridge (Xeon E5)

- 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

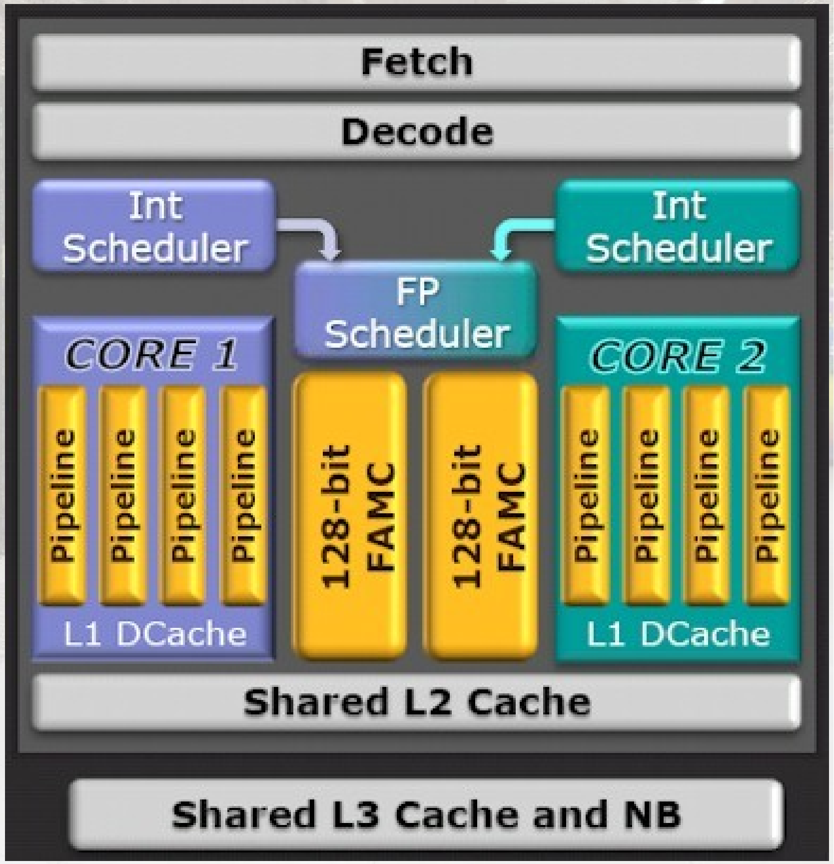
Intel SandyBridge

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 25 people
- Created 1989 when Linköping University purchased a Cray XMP in collaboration with SAAB.

*) Swedish National Infrastructure for Computing

Major Partners and Funding Organisations

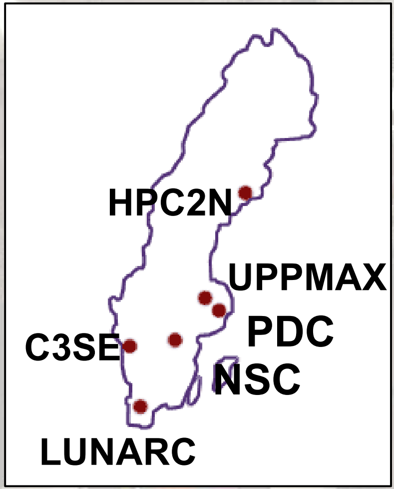
www.snic.vr.se

Swedish National Infrastructure for Computing

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

Provide funding for hardware and support personel for academic projects

Hosted by the Swedish Science Council



www.smhi.se

Swedish Meteorological and Hydrological Institute

Manage and develop information on weather, water and climate that provides knowledge and advanced decision-making data for public services, the private sector and the general public.

www.saabgroup.se

Swedish Aeroplane AB

Products, services and solutions from military defence to civil security.

Mission

- Design, procure and install and maintain computing and storage resources at large scale.
- Provide users with help, training and support to use supercomputers in the most efficient way.
- Manage, monitor, and coordinate systems, facilities, security, and usage.
- Develop and improve large scale computing

Hardware Resources
(approximate numbers)

Computing

- 18000 processor cores
- 170 Teraflops (peak)

Disk Storage

- 3200 drives
(compute servers uncounted)
- 4.6 Petabyte (raw) =
3-3.5 Petabyte user space

Tape Storage

- 2900 slots
- 1700 tapes (LTO4 and LTO5)
- 13 tape drives
- 1.9 Petabyte (raw)

Computing Resources

Academic Projects (SNIC)

Neolith & Kappa
General purpose clusters for academic projects in many different science fields: quantum physics, quantum chemistry, molecular dynamics, climate research, fluid dynamics, and many more.

Neolith

Kappa

Matter
Dedicated cluster for research on new materials for clean energy production, energy storage, emission reduction and nuclear waste disposal.

Matter

Computing Resources
SMHI

Gimle & Bore

Byvind & Bore – Redundant pair of clusters for national weather forecast production at SMHI, including resources for emergencies.

Byvind, located at SMHI

Gimle – Cluster for weather model development and climate research at SMHI

SMHI – Swedish Meteorological and Hydrological Institute (www.smhi.se)

Visit to NSC's computer room

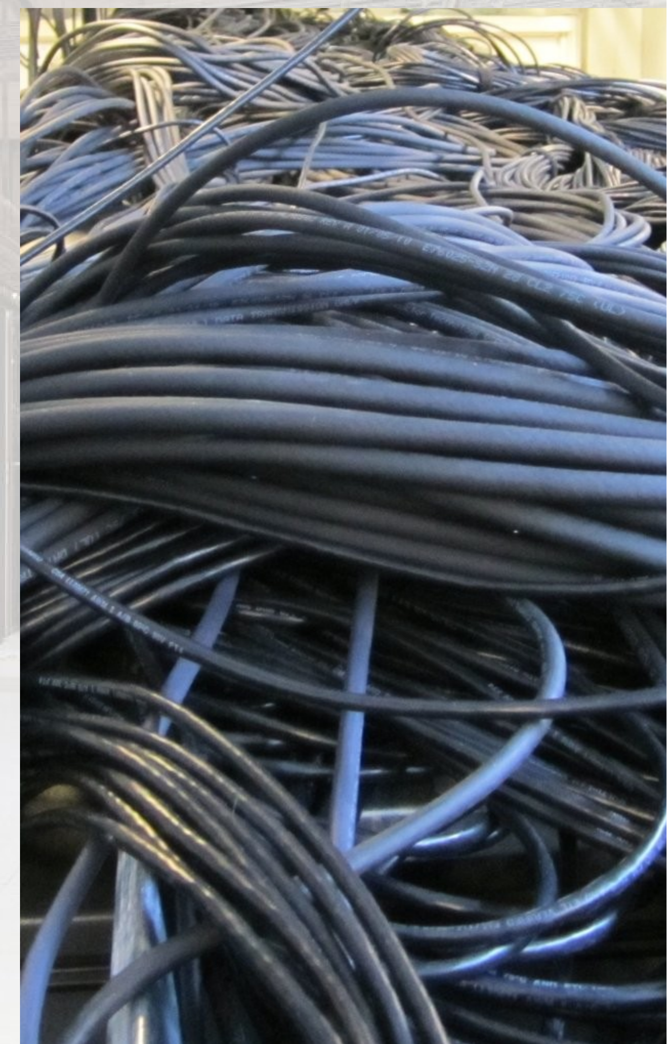
When?

Matter Compute Cluster

- 512 compute servers: HP SL6000
  - 2 x Intel E5520 (8 cores, 2.26 GHz)
  - 36 GiB memory
  - 500 GB disk
- 4 compute servers: HP DL160g6
  - 2 x Intel X5570 (8 cores, 2.93 GHz)
  - 144 GiB memory
  - 4 TB disk
- Central management services
  - Shared with Kappa
  - Private queue manager

Matter Networks

- Application network
  - QDR Infiniband: Voltaire
  - 40 Gb/s
  - Two level fat tree, 1:8 fan-out
- Storage and other services
  - 1 Gb/s Ethernet to servers
  - 10 Gb/s backbone and to storage
- Lights-out network
  - 100 Mb/s
  - Monitoring and power control



Matter Power & Cooling

- 71 16A fuses
  - 140 kW sustained (avg)
- 5 Modular Cooling Systems (MCS)
  - Encapsulated racks (two-by-two)
  - In-row heat-exchangers
  - Increases density and control

Storage Resource for Academic Clusters

GPFS

Neolith

GPFS

Kappa

GPFS

Matter

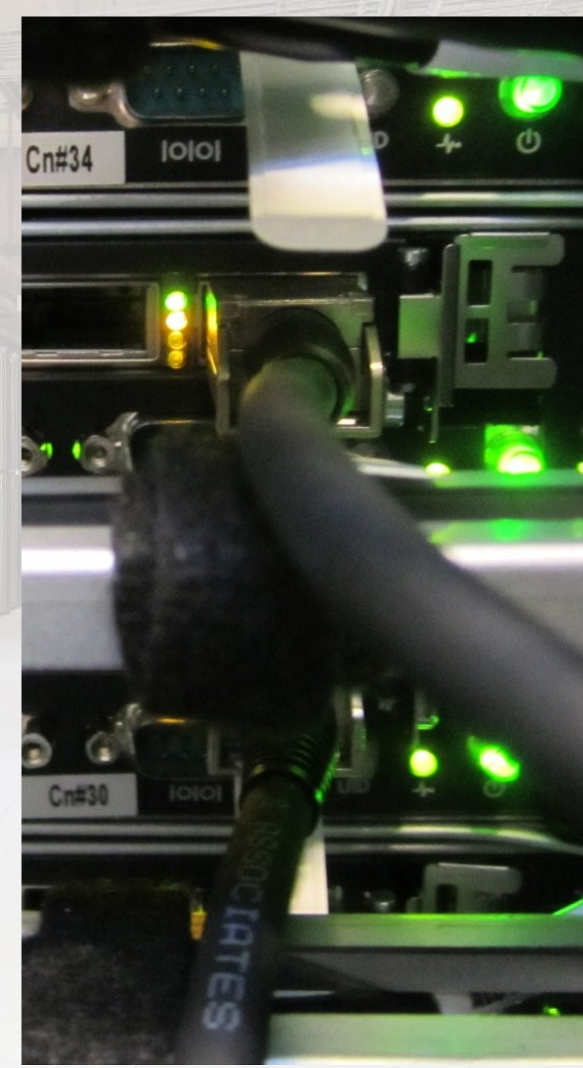
Central Storage
GPFS
16 storage servers
500 TiB RAID6 (504 drives)
approx 1700 clients

Matter Software

- Similar environment on all large clusters at NSC
  - OS: Linux, CentOS distribution
  - Compilers: Intel, PGI, GCC
  - Math Library: Intel MKL
  - MPI: IntelMPI, OpenMPI
  - Job Manager/Scheduler: SLURM
  - Conf. Management: SystemImager, BitTorrent
  - Collectl, nagios, various scripts for management and monitoring
  - Enhanced environment for compiling and starting jobs
- Applications: as needed

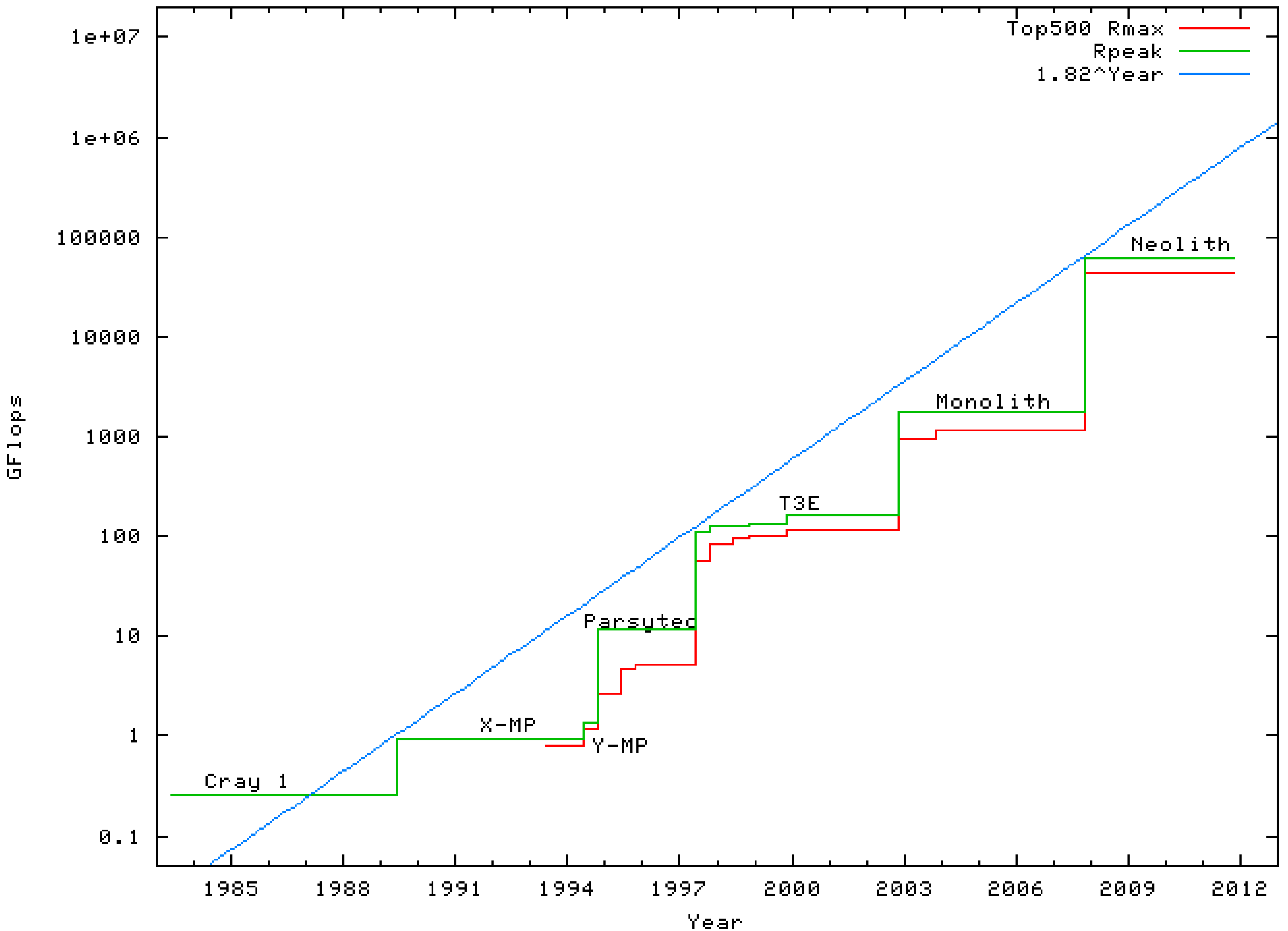
Matter Performance

- Peak Performance: 37.3 Teraflops
- Linpack Perfomance: 34.3 Teraflops (92% of peak)
- Application performance is almost twice as fast as Neolith* on benchmarked applications.



*) Peak: 60 Teraflops, Linpack: 47 Teraflops

Performance of NSC's fastest supercomputers



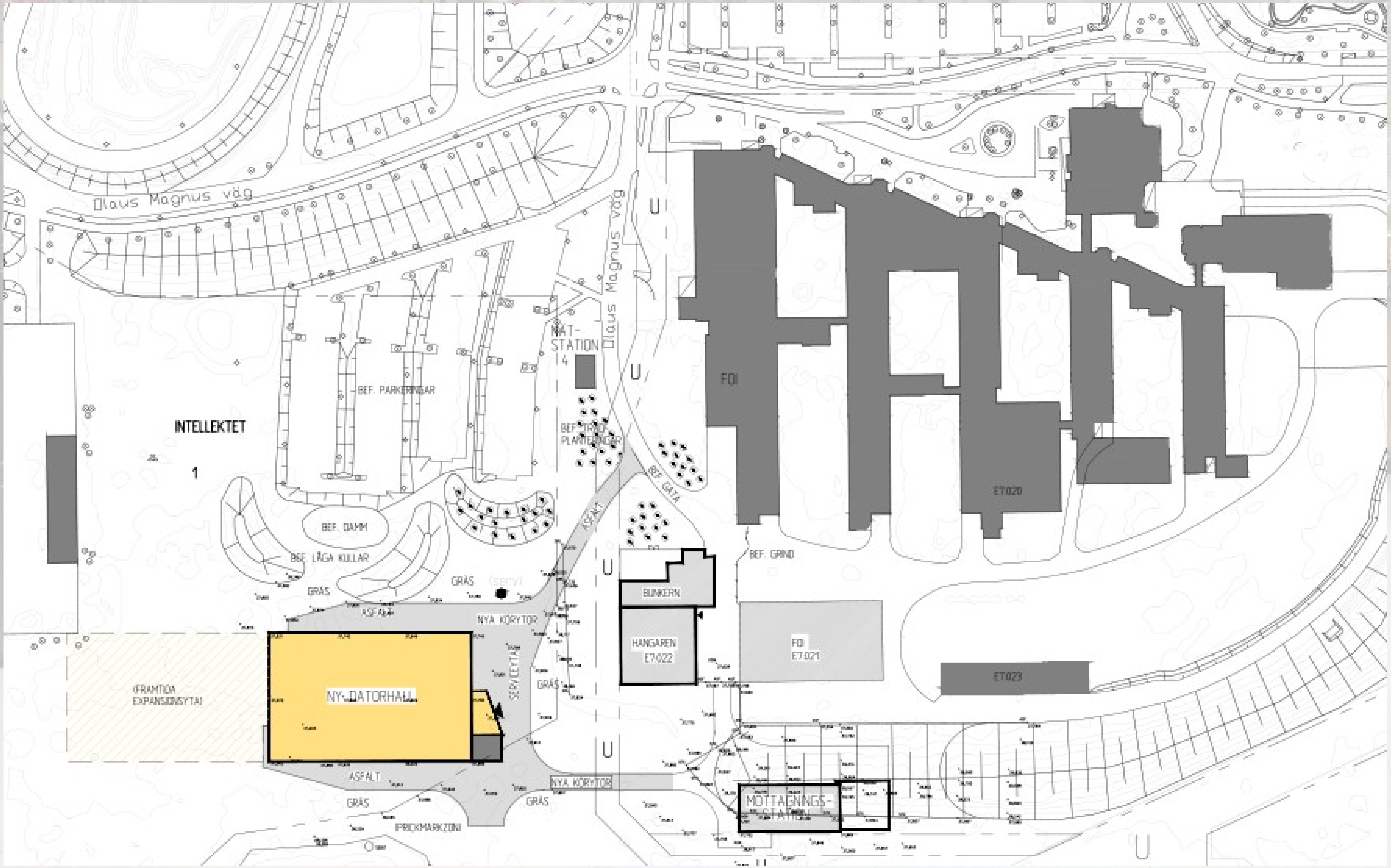
Computer Room Facilities

- Two computer rooms:
  - **Bunkern** (2003) with chiller and air-side economizer
  - **Hangaren** (2007/2009) with district cooling
- Floorspace for IT-equipment: approx 360 m^2 (120 + 240)
- Total used power currently:
  830 kW (24/7/365) = 7.27 GWh/year
  - approx. 40% increase per year during the last 12 years

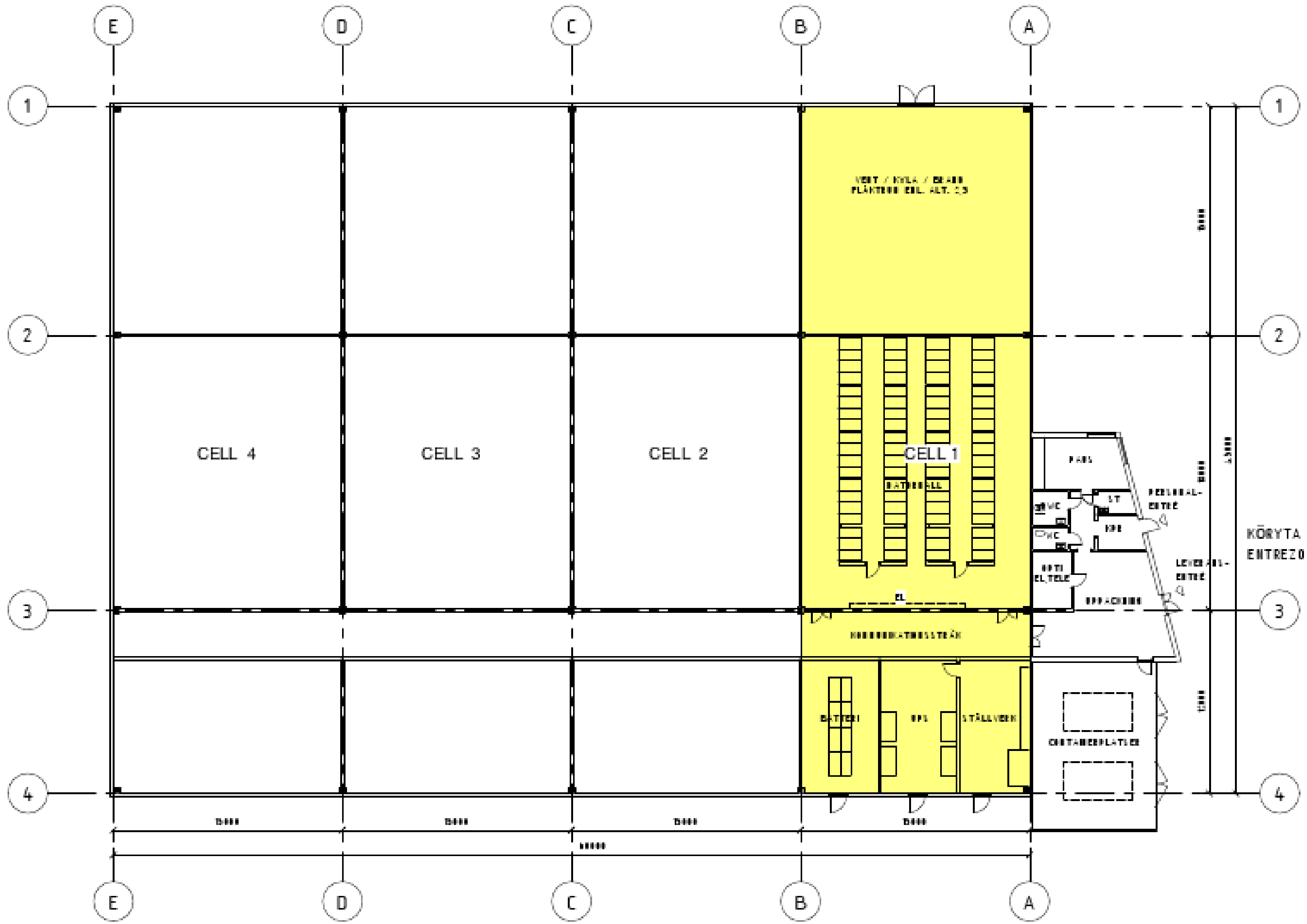
$$\text{PUE} = \frac{\text{Total facility power}}{\text{IT equipment power}}$$

- Power Usage Effectiveness (PUE)
  - Hangaren: 1.17

New Computer Room: Kärnhuset

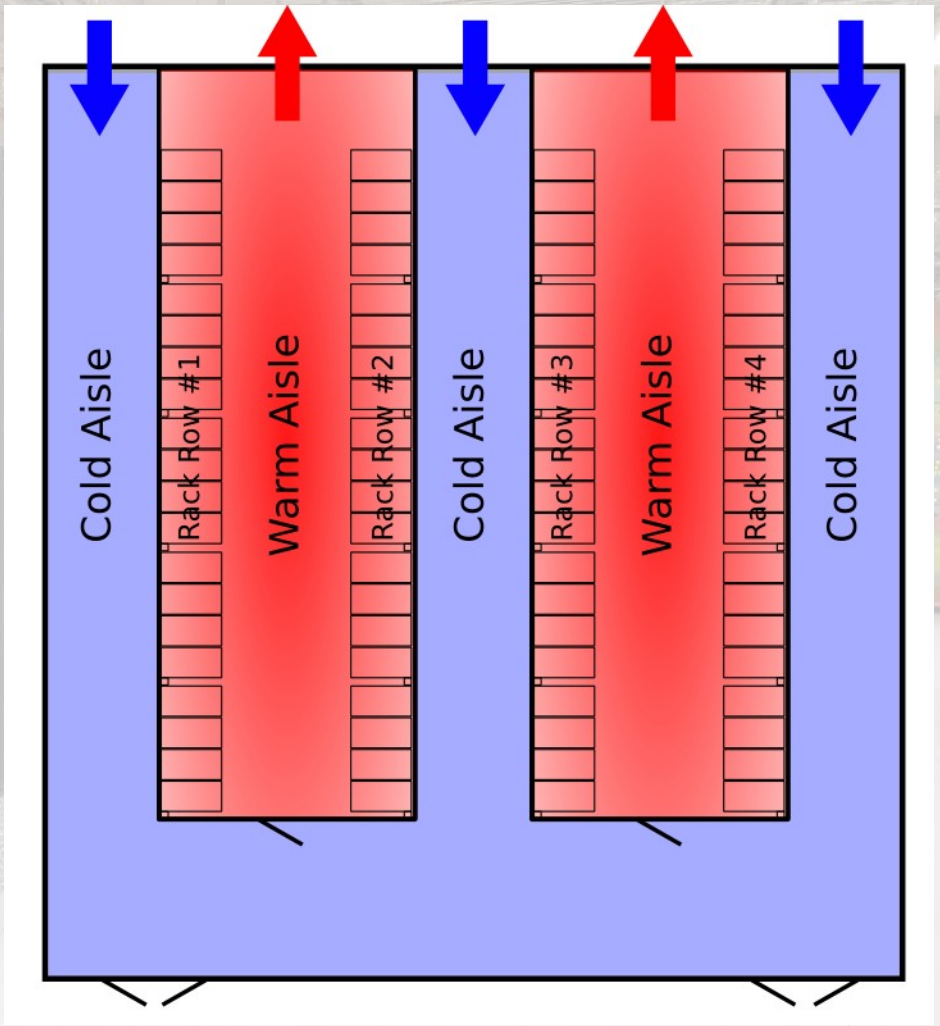


Kärnhuset

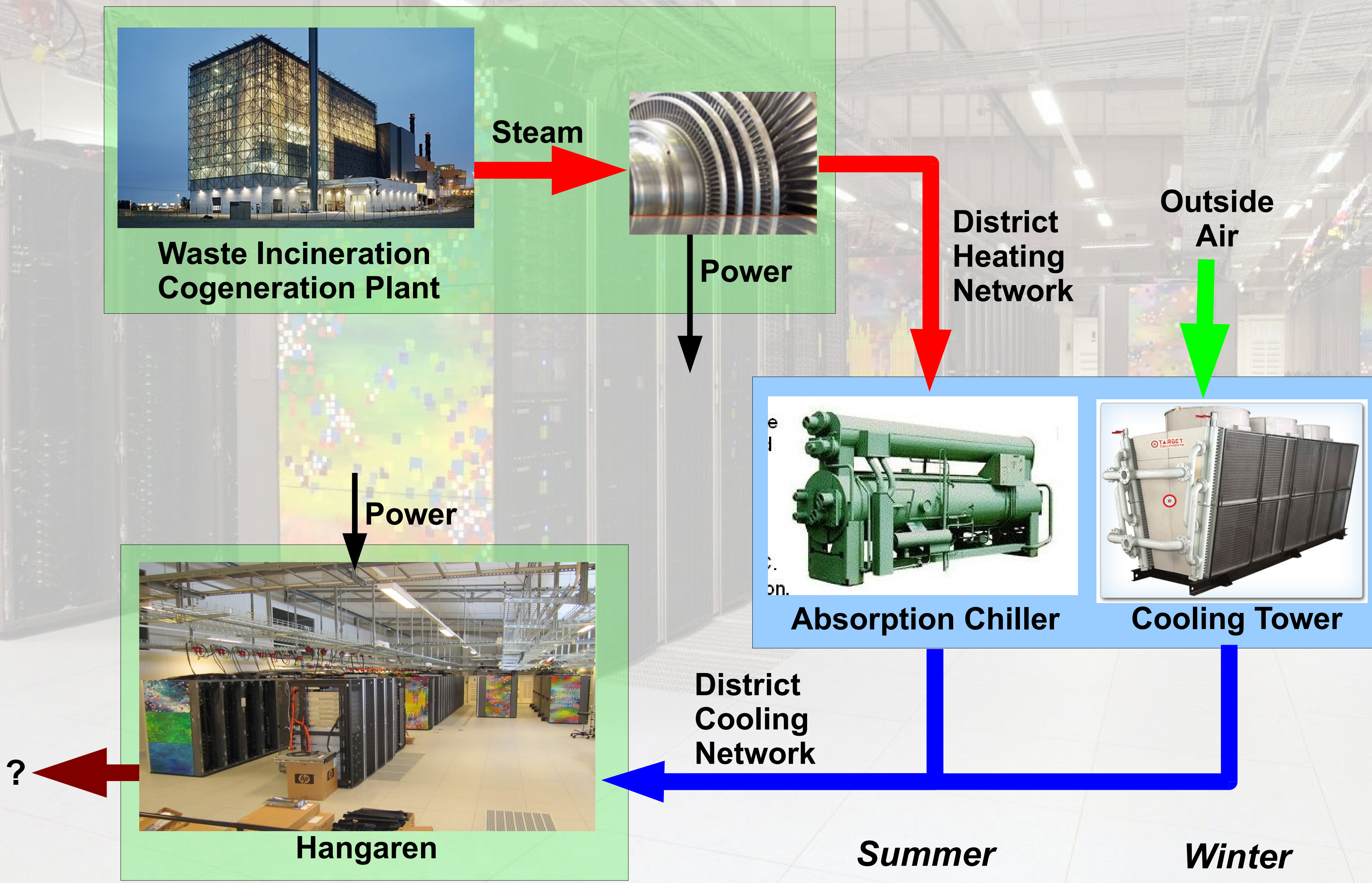


Kärnhuset, Cell #1

- Max 1 MW computer load
- Max 80 racks
- Air cooling
- Aisle separation from the start
- No installation floor
- District cooling
- Ready in 2013 Q2

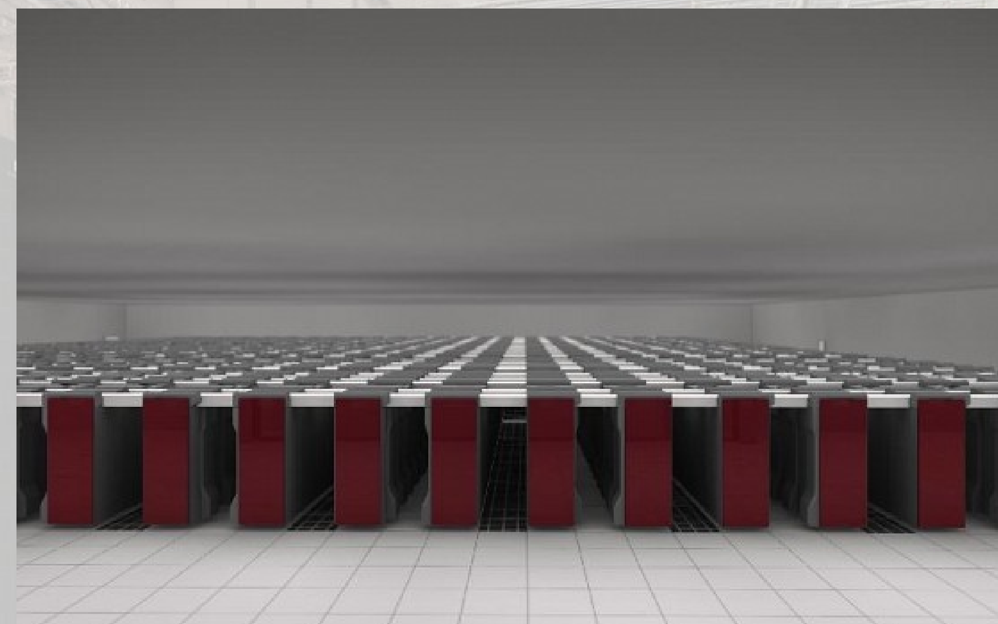


Cooling – Sustainable Campus



No. 1 on Top500: K computer

- 864 cabinets
- 88128 2,0 GHz SPARC VIIIfx
- 705024 cores
- 1,3 PiB memory
- 6 dimensional Tofu network
- Peak: 11,28 petaFLOPS
- Linpack: 10,51 petaFLOPS
- 12,6 MW



Future: exaFLOPS

Targets by DARPA:

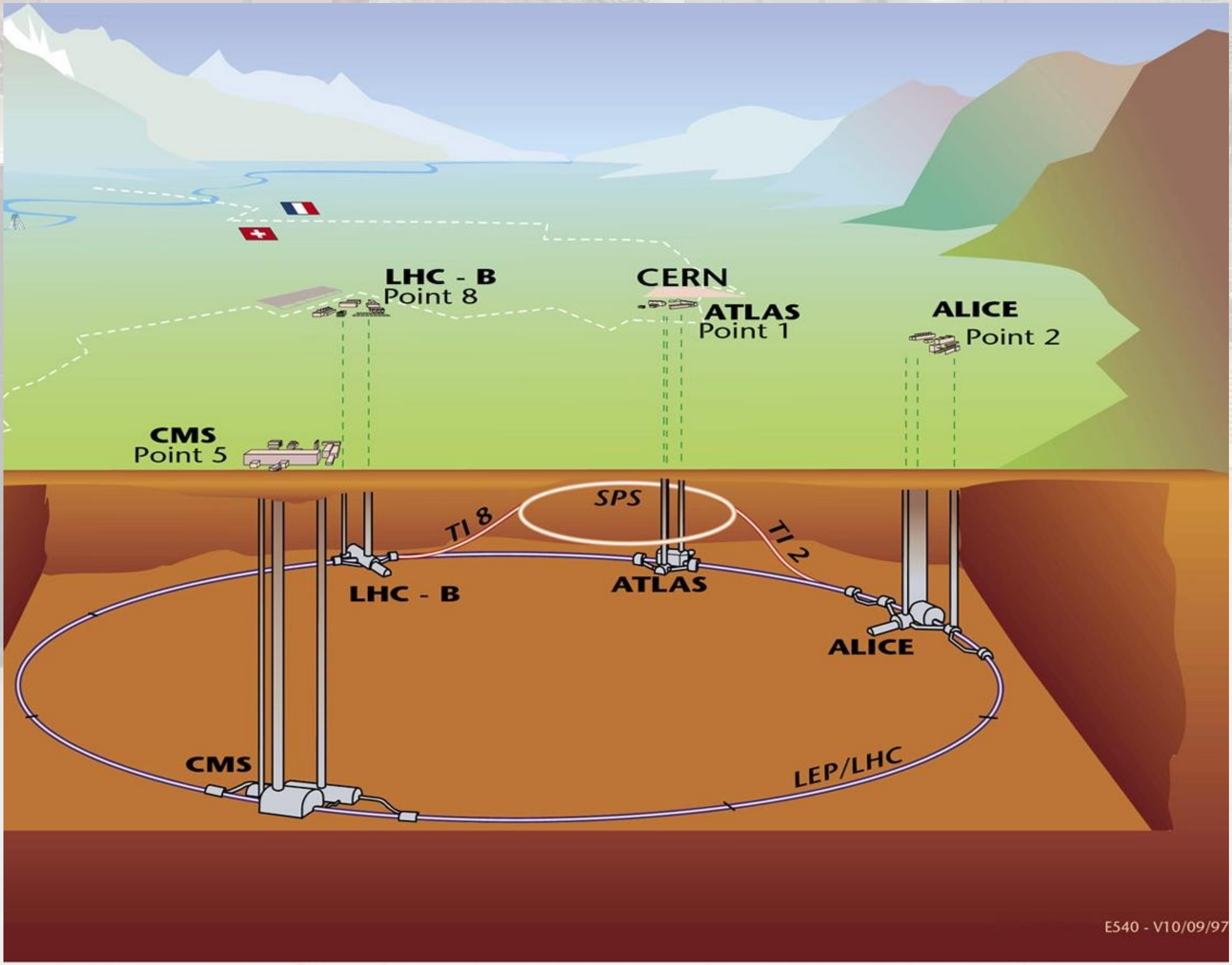
- 2018: 1 exaFLOPS
  - 2008: 1 petaFLOPS, LANL, IBM
  - 1998: 1 teraFLOPS, ASCI Red, Sandia
- 32-64 PiB memory
- ~20 MW
- MTTI: O(1 day)

Applications

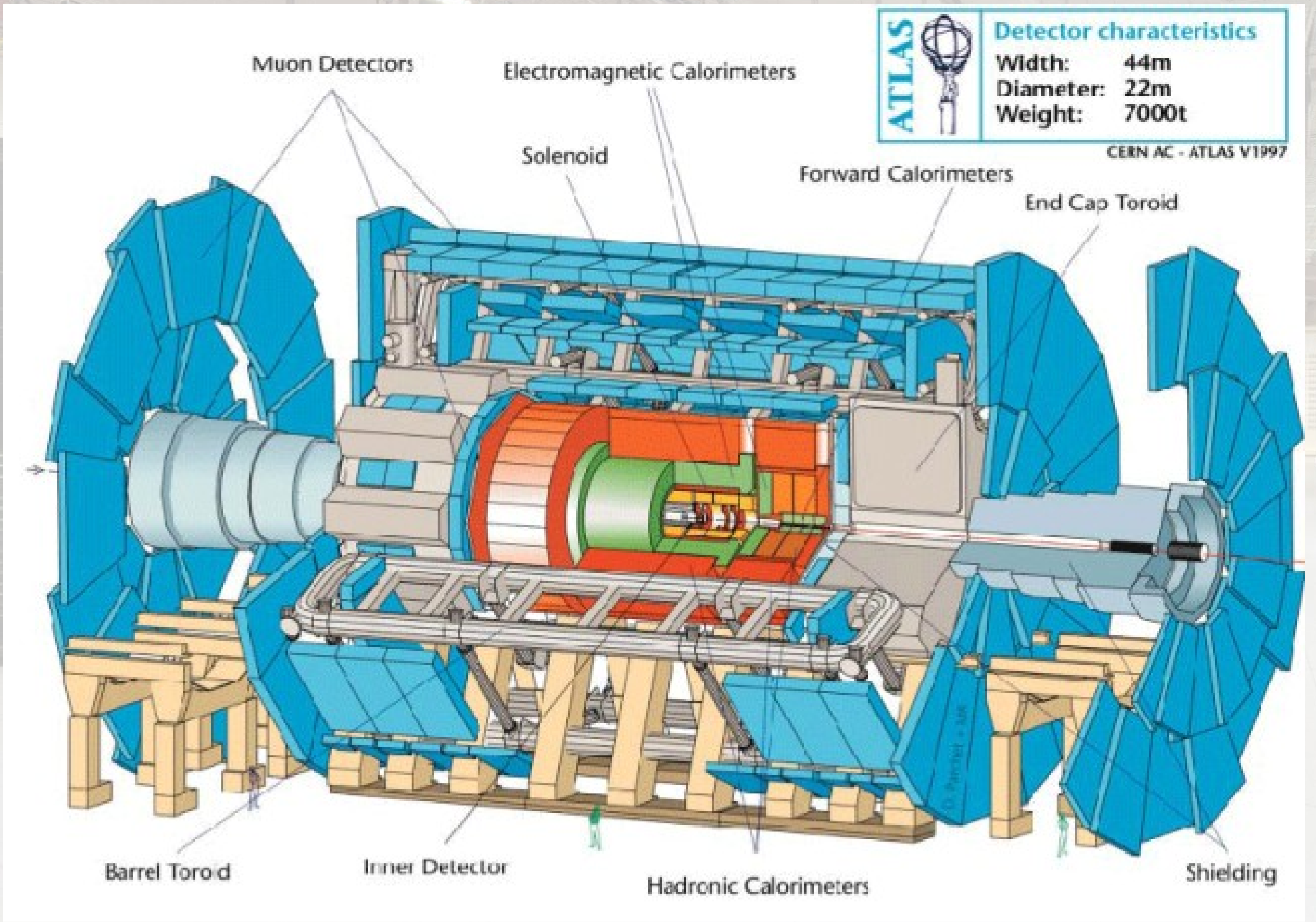
- Climate
  - Extreme weather
  - Carbon, Methane, and Nitrogen cycles
  - CO_2 sequestration
  - Scenario replications, ensembles
  - Increase time scale
- Computational Fluid Dynamics
  - Design of aircrafts, vehicles, submarines
  - Combustion, Turbulence
- Advanced materials
  - Solar cells
  - Fuel 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

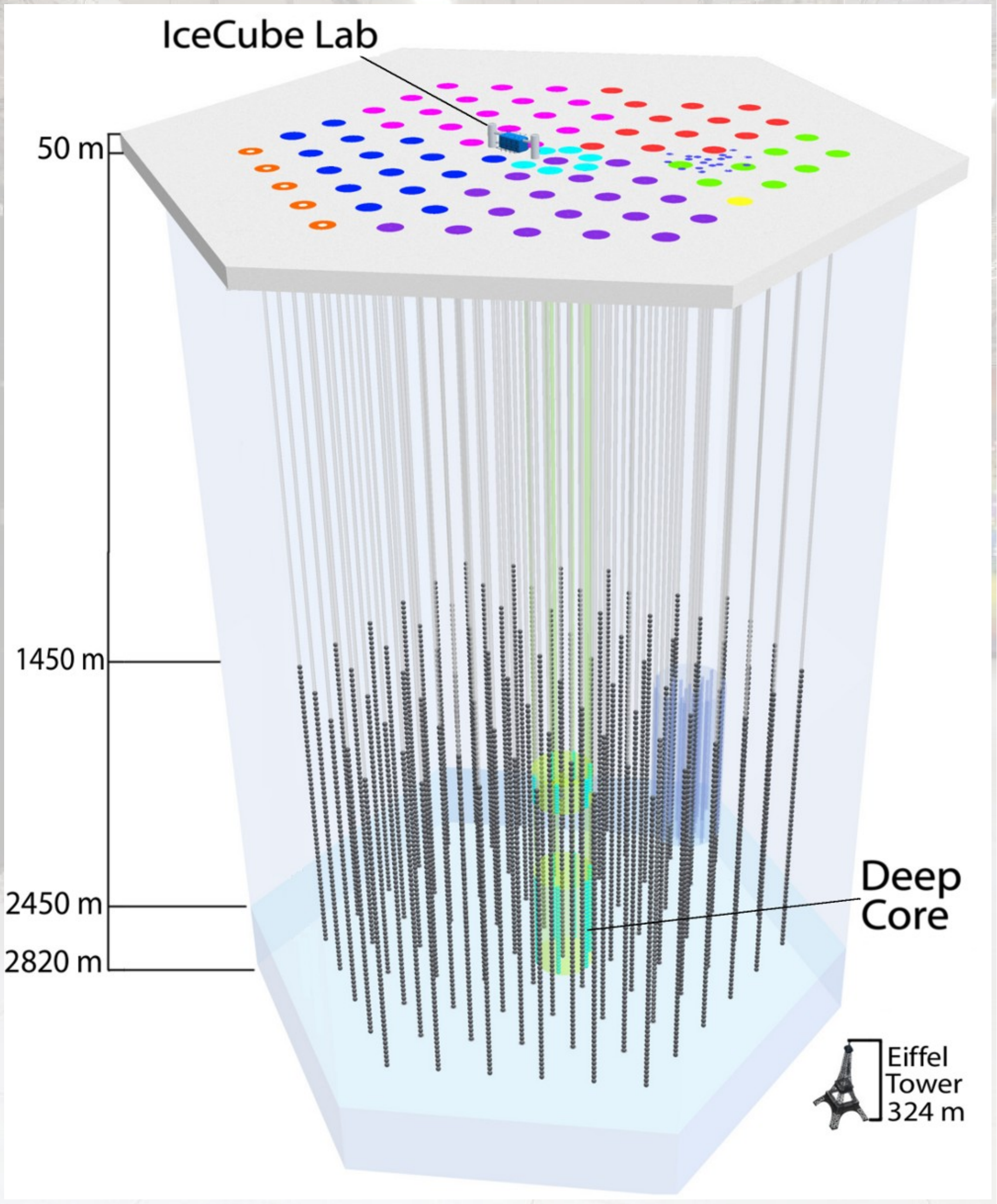


ATLAS Detector

IceCube - Neutrino observatoriet

Klas Hultqvist, Stockholms universitet

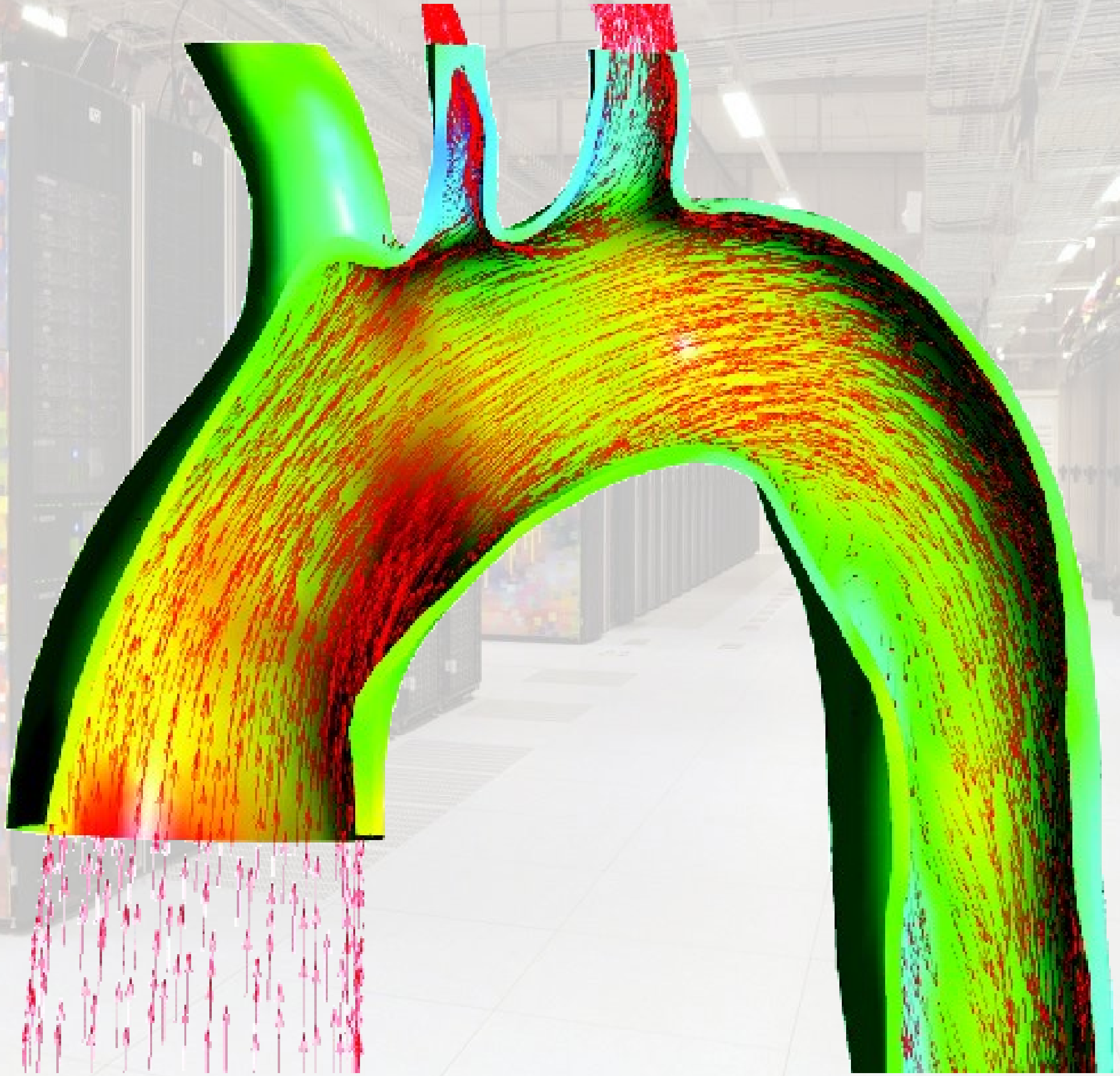
- Detektor vid sydpolen
- 86 vertikala band
- 5160 optiska moduler i 1 km^3 is
- Detekterar Čerenkov strålning från sekundära partiklar neutrinos → muons
- 2000 händelser eller 10 MByte data per sekund
- 100 TByte data per år



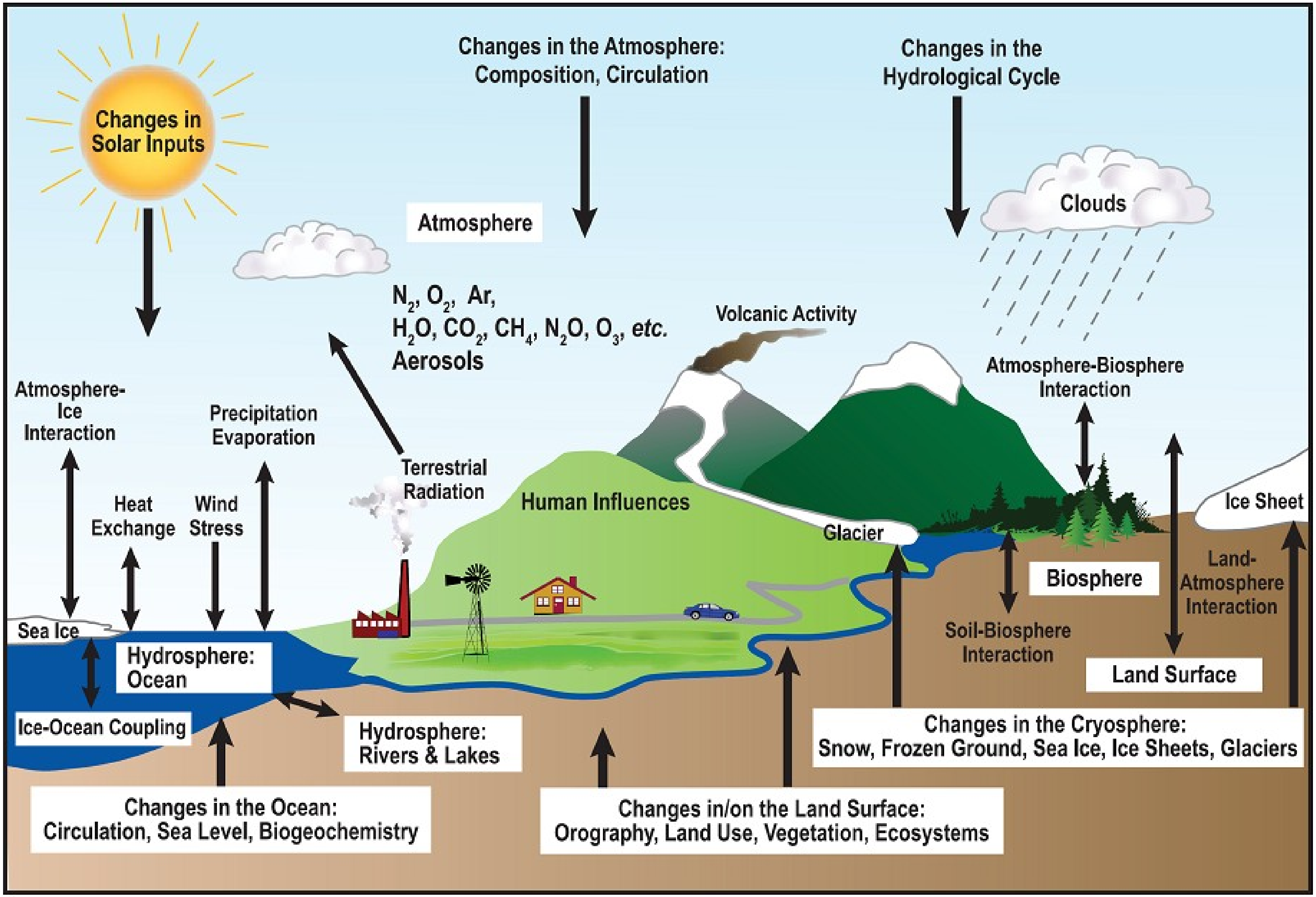
Simulering av blodflöde i aortan

Matts Karlsson, Linköpings universitet

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



Climate Simulation

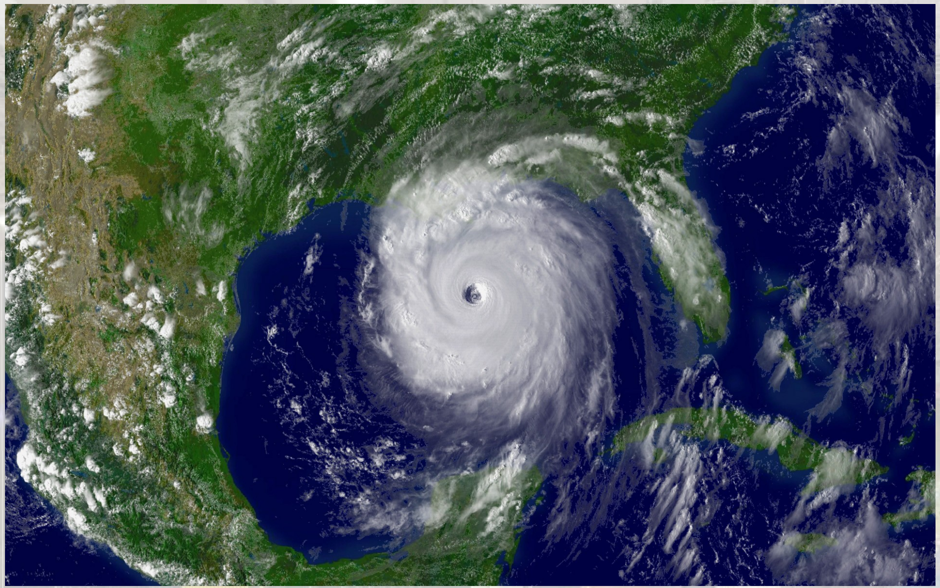


FAQ 1.2, Figure 1. *Schematic view of the components of the climate system, their processes and interactions.*

Numerical Weather Prediction

Challenges

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

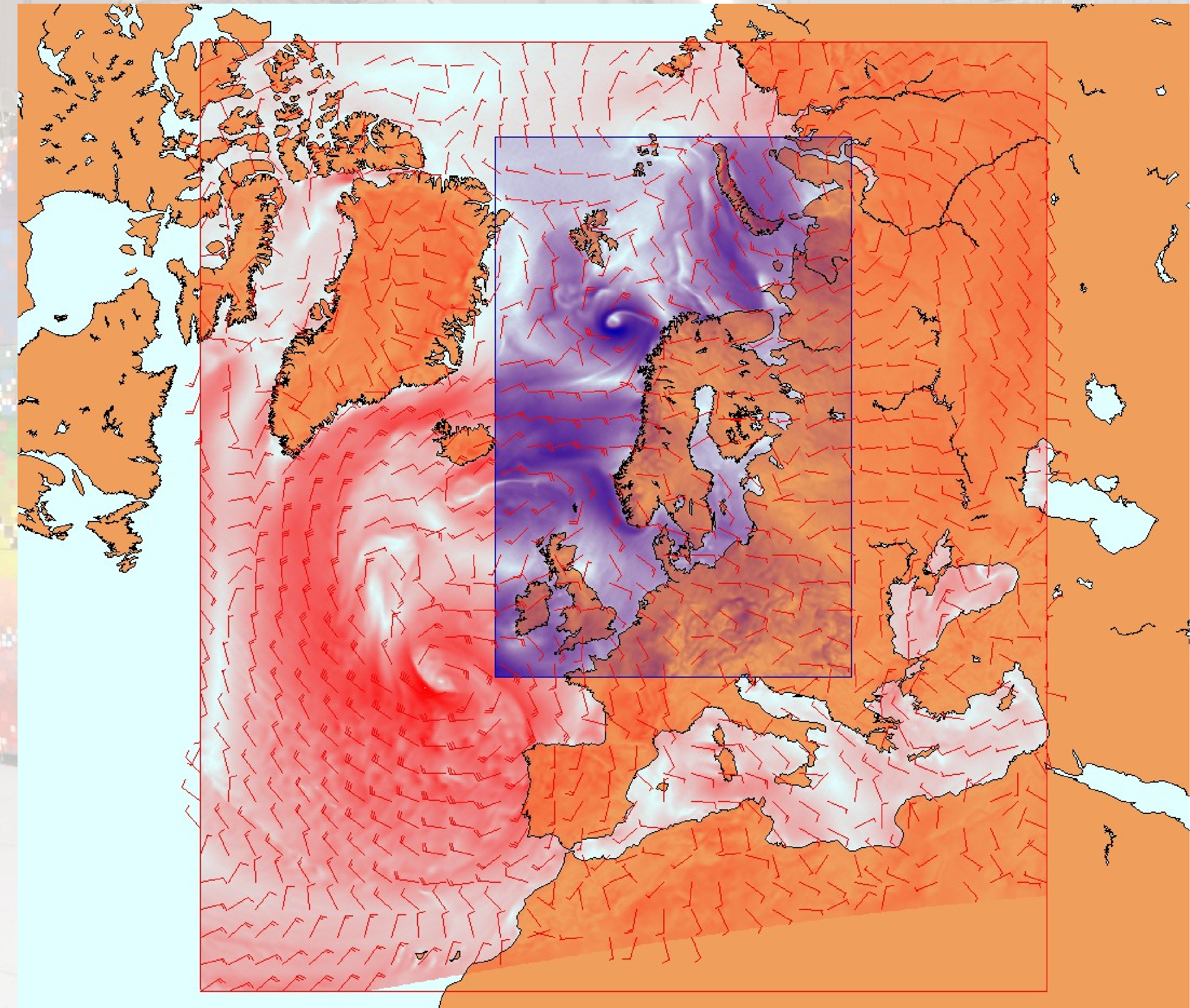


Hurricane Katarina, 2005

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




Cn#34
Cn#30