GPU - GreenGrid - VRE

Harry Enke,

Astrophysical Institute Potsdam

AIP

GPU - GreenGrid - VRE

- **Research topics & areas with special hardware**
  - **Numerical Astrophysics with GRAPE/FPGA/GPU**
  - **Data Management for huge astronomical data archives**
  - **Robotic Telescopes**
- **Collaborative use of resources**
  - **Green Grid**
  - **Virtual Research Environment**

GPU - GreenGrid - VRE

The slides on
GPU, FPGA and GRAPE
are mostly courtesy of
Rainer Spurzem and collaborators

(R. Spurzem, National Astronomical Observatories (NAOC)
Chinese Academy of Sciences (CAS), Beijing, China + ZAH)

Research Areas

Astrophysical / Physical Science Simulations

Stellar Dynamics, Galactic Nuclei, Galaxies, Numerical Relativity, ...

Data Processing

Data Processing pipelines, Analysis pipelines

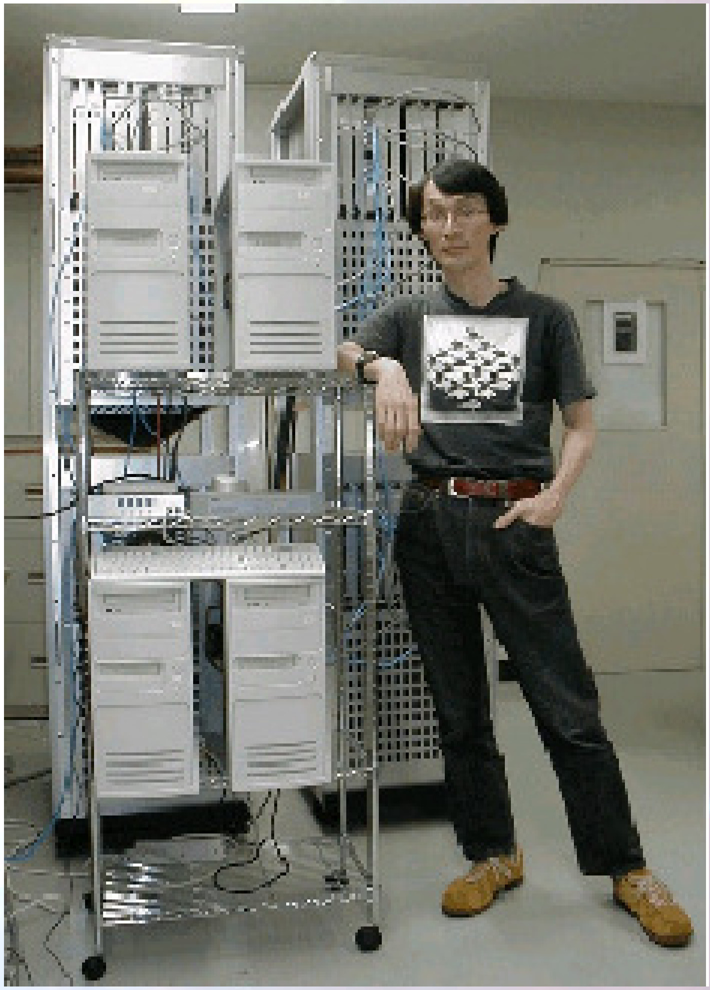
- Computing - development of -
- Libraries, Tools
- Software-framework
- Hardware test-beds
- Coupled Hardware/Software turnkey solutions

GRAvity PipE (GRAPE)

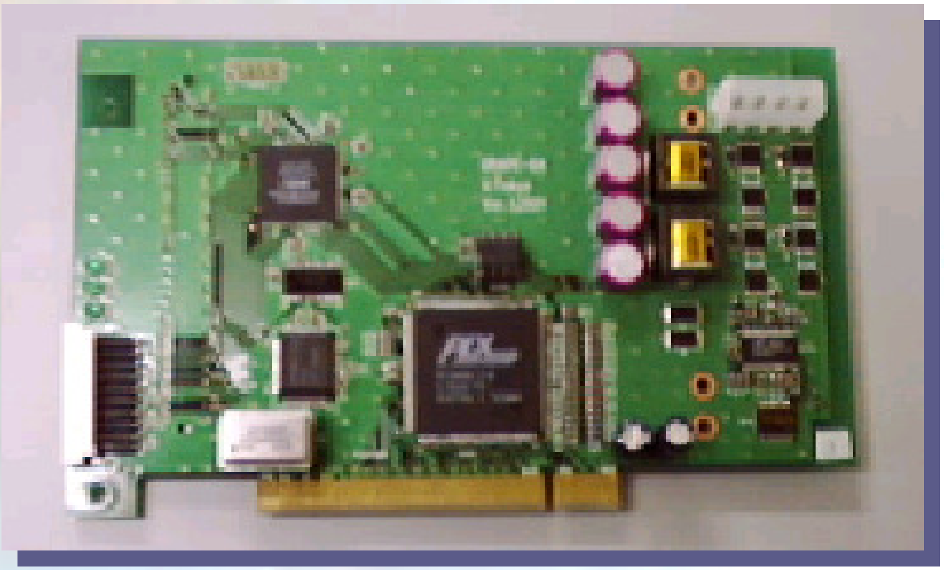
GRAPE-6 Gravity/Coulomb Part

- G6 Chip: 0.25μ 2MGate ASIC, 6 Pipelines
- at 90MHz, 31Gflops/chip
- 48Tflops full system (March 2002)
- Plan up to 72Tflops full system (in 2002)
- Installed in Cambridge, Marseille, Drexel, Amsterdam, New York (AMNH), Mitaka (NAO), Tokyo, etc..

New Jersey, Indiana, Heidelberg



GRAvity PipE (GRAPE)



GRAPE6a PCI board

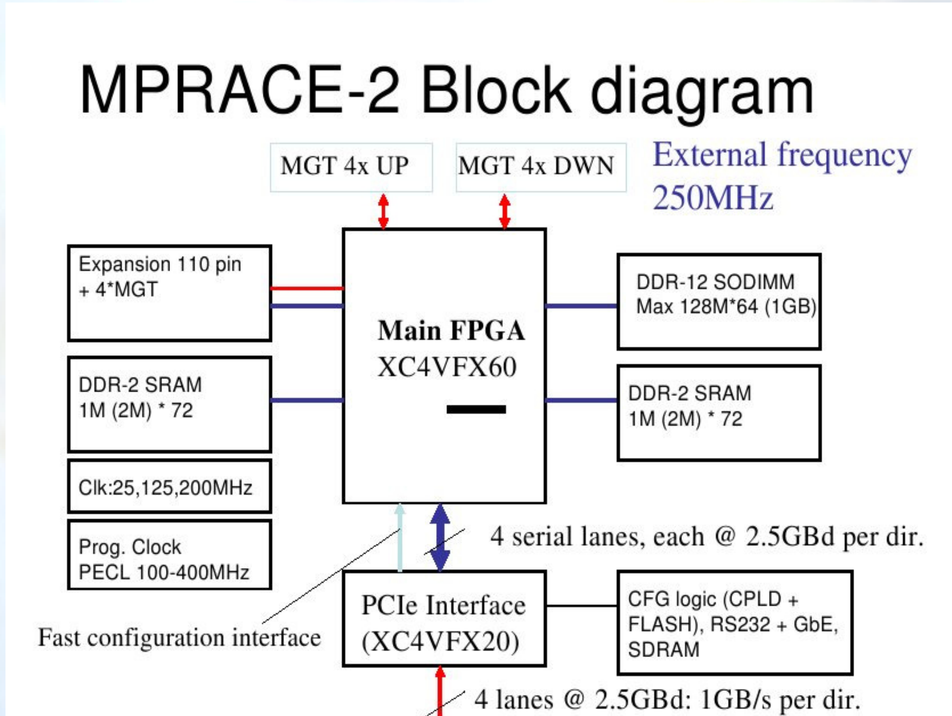
GRAPE6a - PCI Board for PC-Clusters

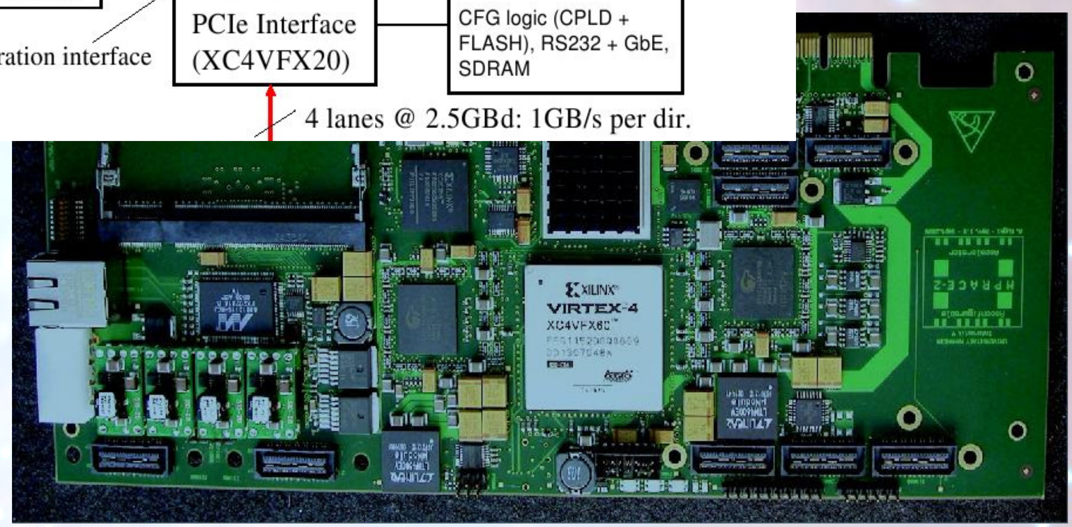
~128 Gflops for a price ~5K USD;
Memory for up to 128K particles

Field Programmable Gate Arrays (FPGA)

Computer Engineering
R. Männer
G. Marcus
A. Kugel

Univ. Heidelberg
(Mannheim)

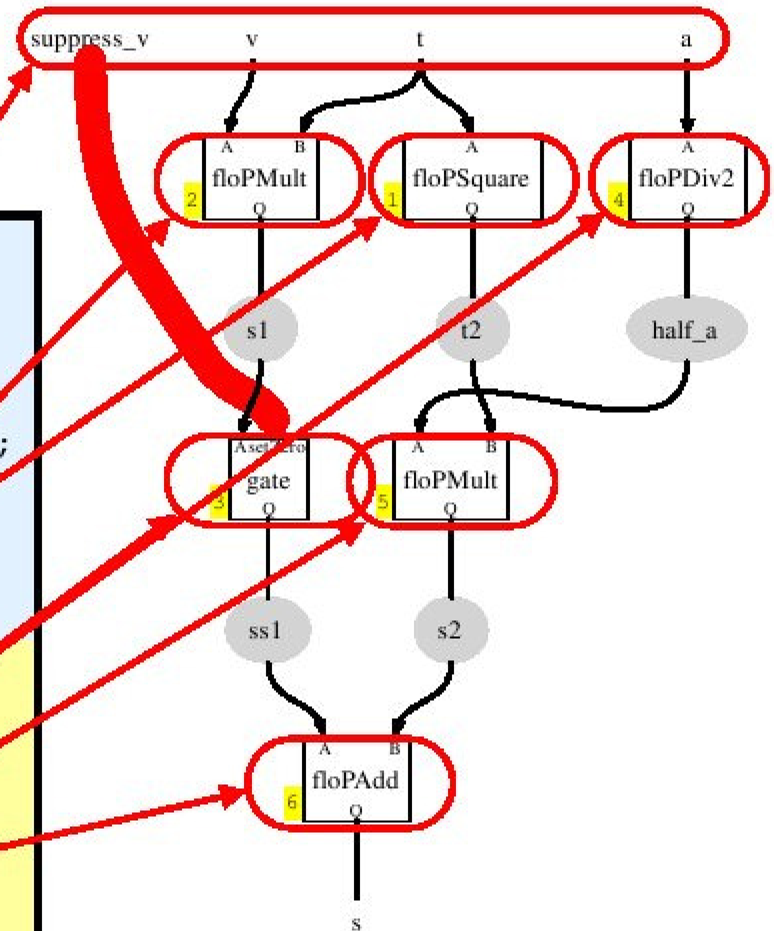




Field Programmable Gate Arrays (FPGA): Pipeline Generation

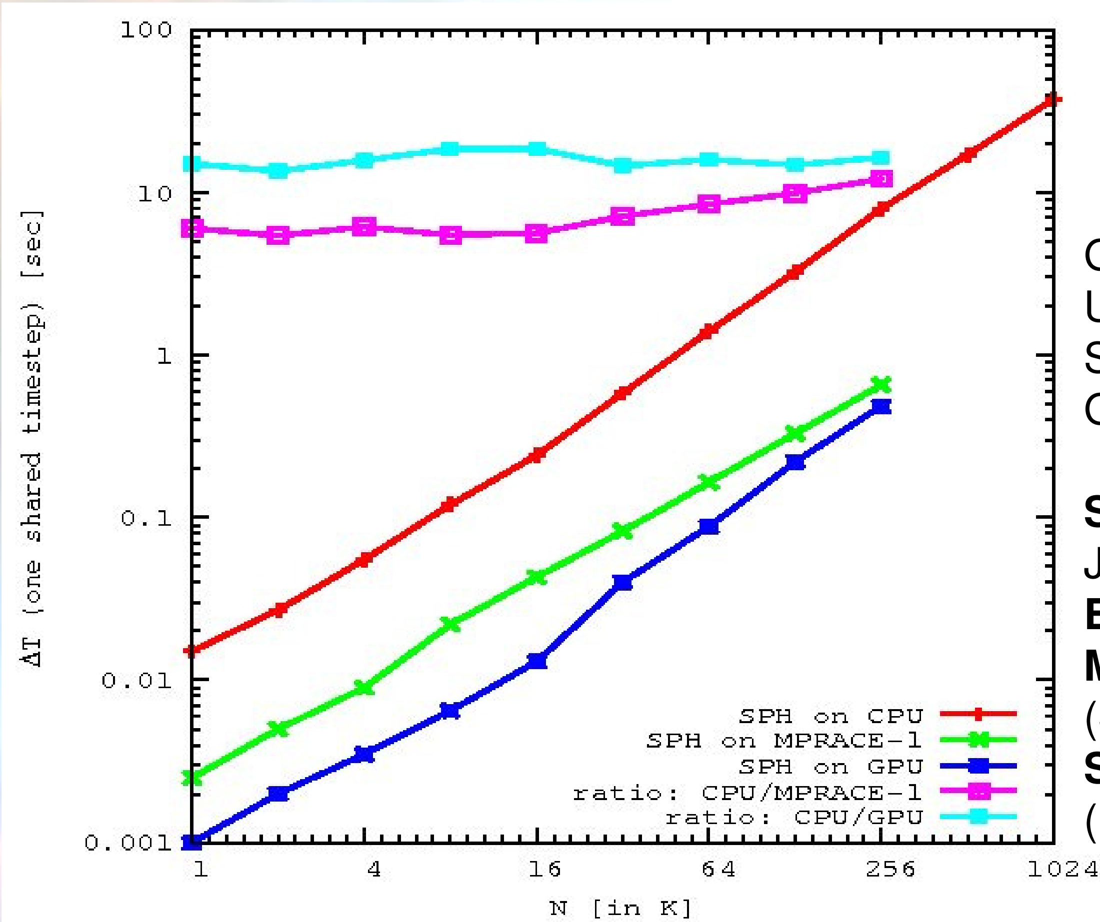
$$s = \begin{cases} \frac{1}{2} a\, t^2 + v\, t & : \quad \text{suppress\_v} = 0 \\ \frac{1}{2} a\, t^2 & : \quad \text{suppress\_v} = 1 \end{cases}$$

```
entity distance;
clock clk;

# parameters
floPValDef fpDef(signifLength=>24,
 expLength=>8,useSign=>1,useIsZero=>0);

# inputs
signal (suppress_v);
floPVal (v,a,t)(fpDef);

# calculate
t2 = <floPSquare> t;
s1 = v <floPMult> t;
ss1 = gated(s1,suppress_v);
half_a = <floPDiv2> a;
s2 = half_a <floPMult> t2;
s = ss1 <floPAdd> s2;
```



Arrays (FPGA):
SPH Application: Power Consumption / Scalability

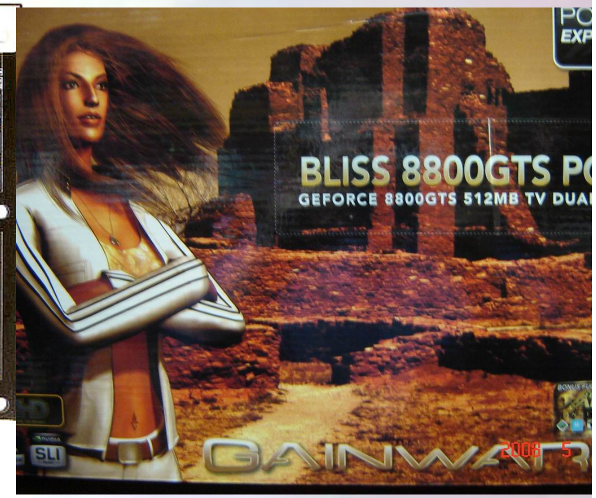
FPGA versus GPU for SPH:



GeForce 8800 GTX (NVIDIA)
Using CUDA Library
Special Interfaces and API from GRACE project ported.

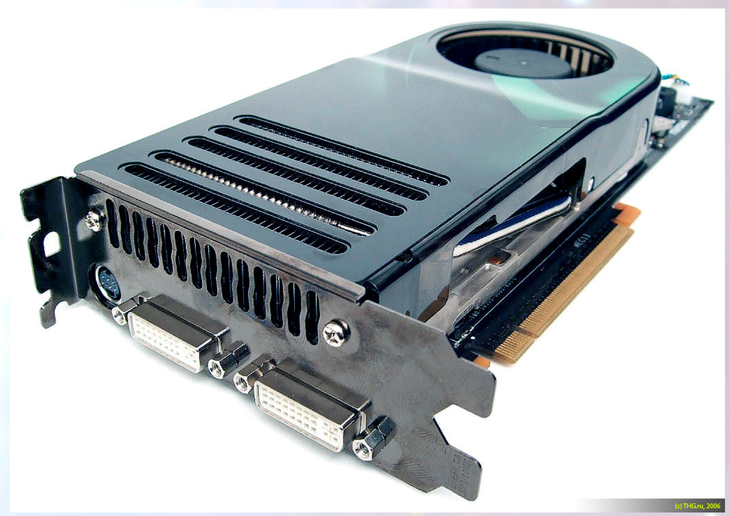
Spurzem et al. 2007,
Jl.Phys.Conf.Ser.
Berczik et al. 2008,
Marcus et al. 2008
(SPHERIC)
Spurzem et al. 2009
(ISC 09 Procs., Springer)

GPU : Graphic Processing Unit



2007…
GeForce 8800 GTX, 128 Stream Proc., 768 MB
GeForce 8800 GTS, 128 Stream Proc., 512 MB
GeForce 8800 GT, 112 Stream Proc., 512 MB

2008…
GeForce 9800 GTX, 128 Stream Proc., 512 MB
GeForce 9800 GX2, 256 Stream Proc., 1 GB
GeForce 9800 GT, 64 Stream Proc., 512 MB



http://www.nvidia.com

GPU Computing

General Purpose GPU Supercomputing (GPGPU)

http://www.nvidia.com
http://www.astrogpu.org
http://gpgpu.org

Key Features

GPU

- Number of processor cores: 240
- Processor core clock: 1.296 GHz
- Voltage: 1.1875 V
- Package size: 45.0 mm × 45.0 mm 2236-pin flip-chip ball grid array (FCBGA)

Board

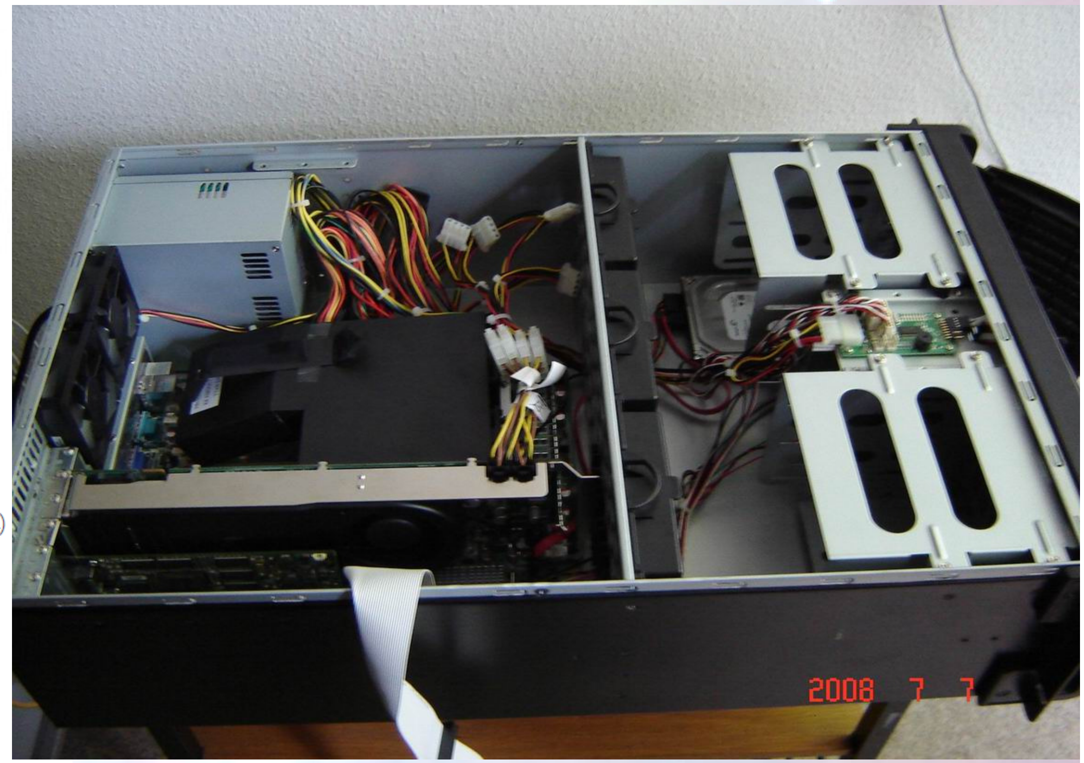
- Fourteen layer printed circuit board (PCB)
- PCI Express 2.0 ×16 system interface
- Physical dimensions: 4.376 inches × 10.50 inches, dual slot
- Board power dissipation: 187.8 W

External Connectors

- None

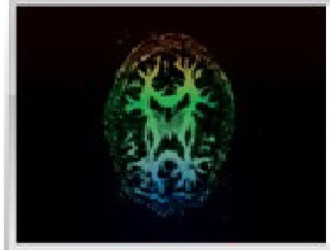
Internal Connectors and Headers

- One 6-pin PCI Express power connector
- One 8-pin PCI Express power connector
- 4-pin fan connector



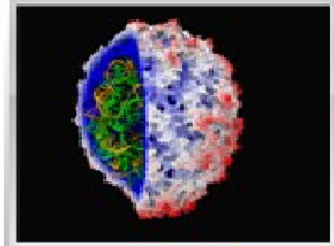
NVIDIA Tesla C1060
in kolob cluster Heidelberg Univ.

Speedups using GPU vs. CPU



146X

Interactive visualization of volumetric white matter connectivity[1]

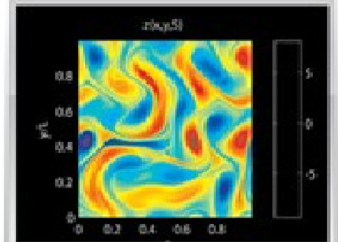


36X

Ionic placement for molecular dynamics simulation on GPU[2]

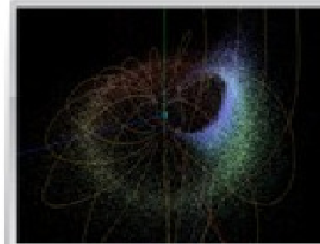
18X

Transcoding HD video stream to H.264 for portable video[3]



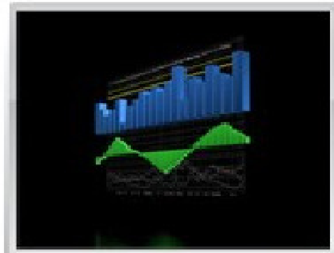
17X

Simulation in Matlab using .mex file CUDA function[4]



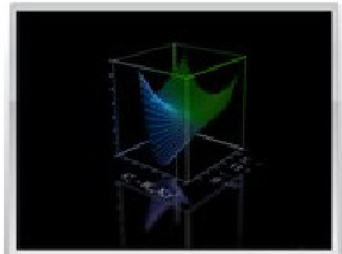
100X

Astrophysics N-body simulation[5]



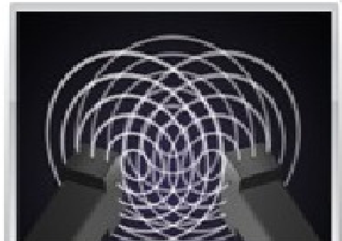
149X

Financial simulation of LIBOR model with swaptions[6]



47X

GLAME@lab: M-script API for linear Algebra operations on

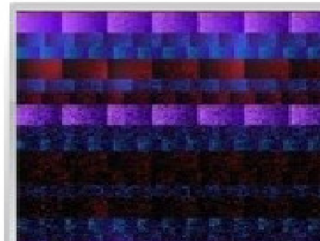


20X

Ultrasound medical imaging for cancer

24X

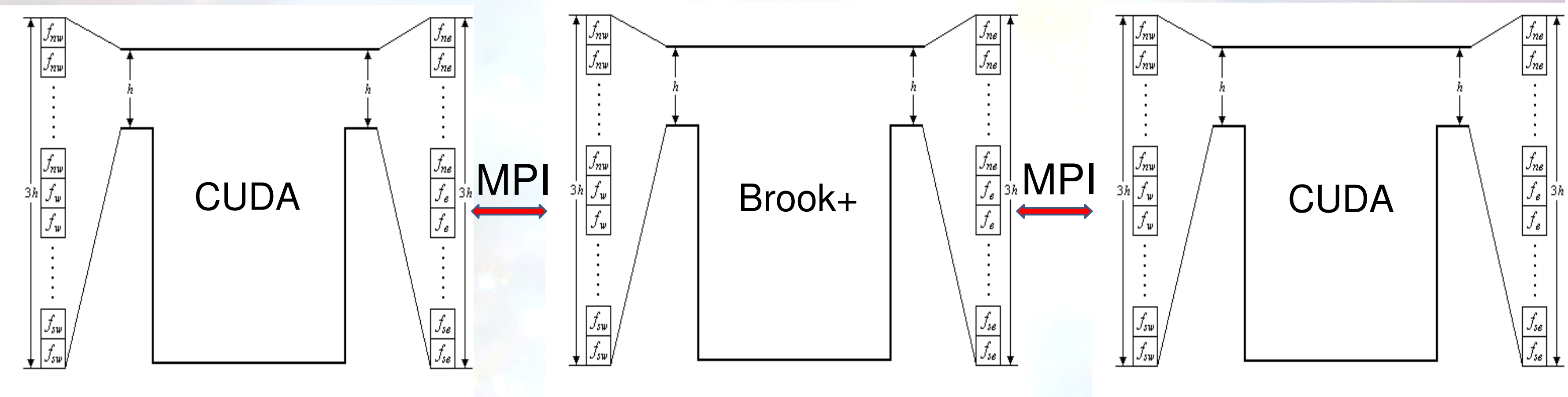
Highly optimized object oriented molecular

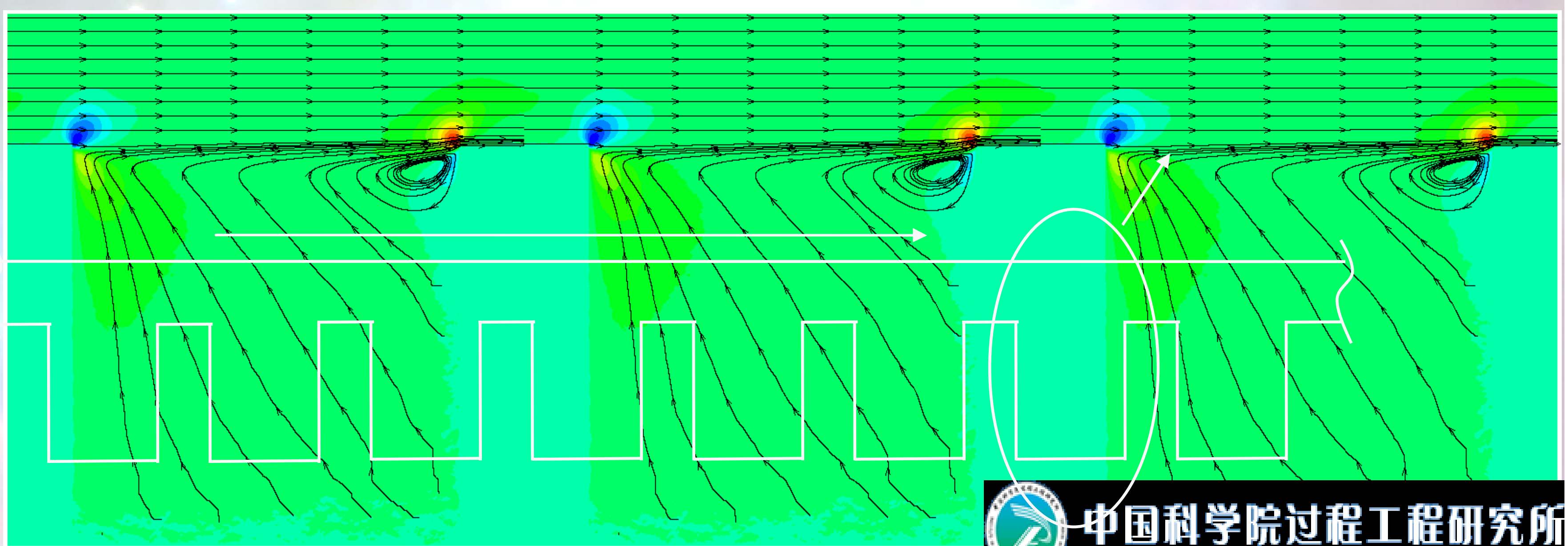


30X

Cmatch exact string matching - find similar proteins & gene

Example: Couette-Cavity flow

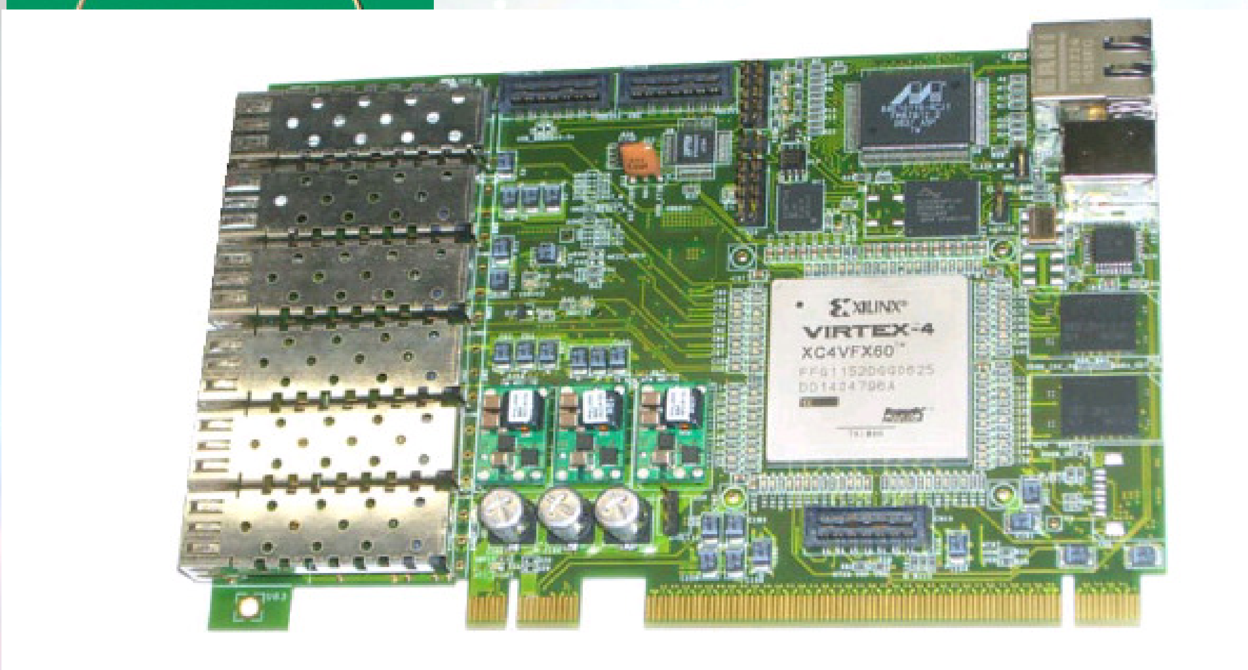




Further Hardware - Speedups

Custom Hypertransport Based Interconnect using FPGA Interface Cards, at the Univ. of Heidelberg, partnership with AMD

U. Brüning and collaborators



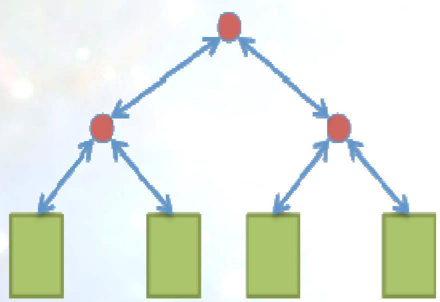
HTX Board
sub microsec latency
EXTOLL Interconnect
switchless

<u>shared / distribute memory boundary blurs!</u>

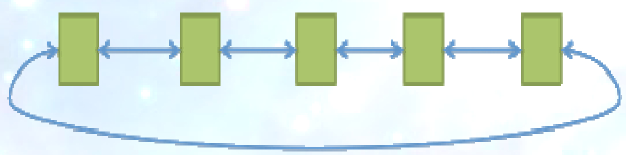
Further Hardware - Speedups

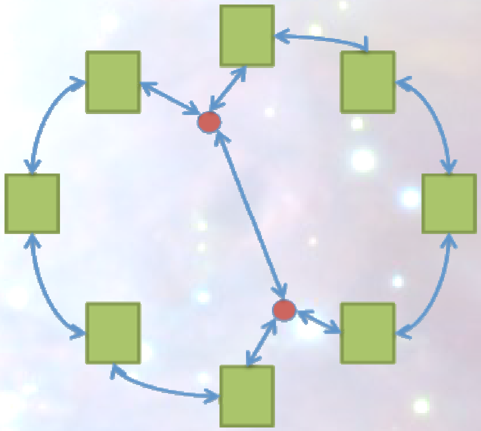
Flexible Switch-Free Adaptive Network Architecture Using EXTOLL High-Speed Interconnect

Hierarchical Tree



Token Ring



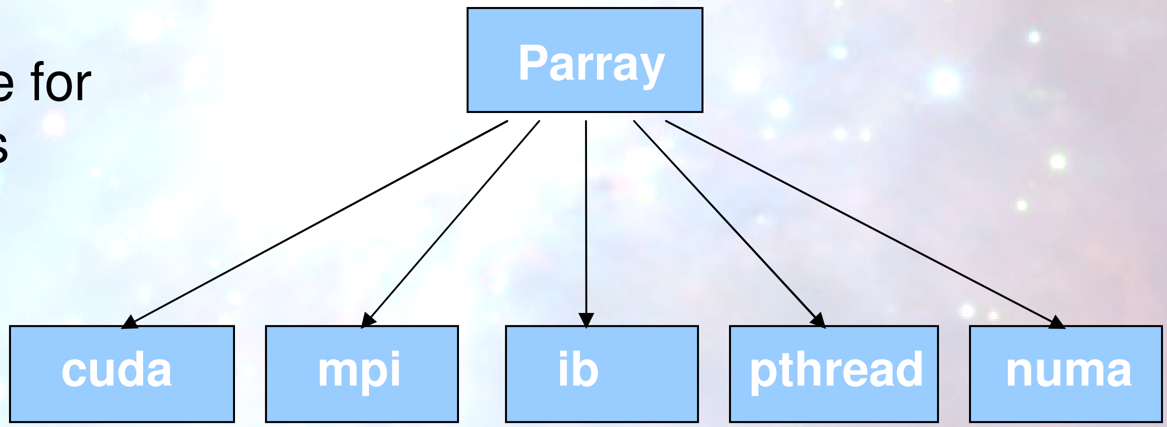


Token Ring with Shortcut

Code Interfaces

- Full Gravity for GPU (phiGPU with Nitadori)
- SPH for GPU
- SPH for FPGA
- NBODY6++GPU (in prep.)
- Tree for GPU (with Hamada, Nitadori)
- GADGET (in prep.)
- More in coop. (FFT, Data, Molecular Dynamics, Numerical Relativity)

- Abstract Interface for Parallel Arrays



Hardware for astronomical Databases: GrayWulf-Architecture*

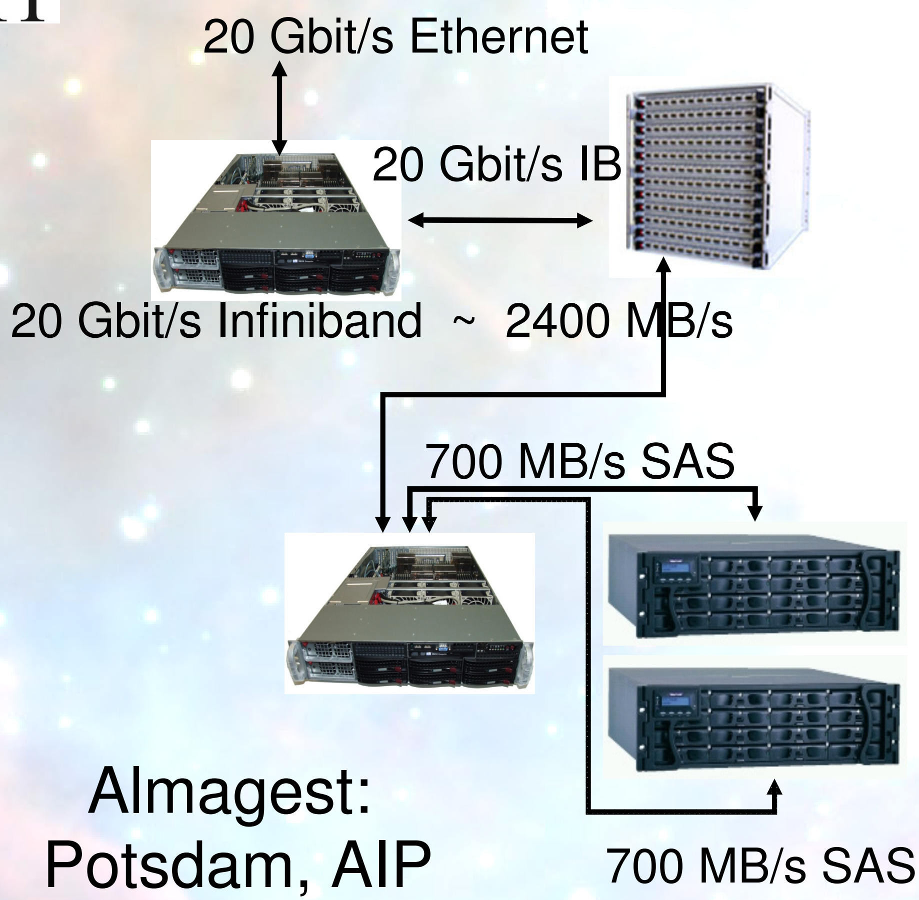
GrayWulf architecture is based on commodity hardware

Principles:

- Each server has a separate raid controller (Areca) for each JBOD
- JBOD to raid controller has 4x3 SAS lanes
- SAS harddisks (Seagate, 1TB) avoid protocol overhead of ~20% compared to SAS/SATA
  - (most current raid systems feature SAS interconnect and SATA harddisks)

*A. Szalay et al., GrayWulf: Scalable Clustered Architecture for Data Intensive Computing

Hardware for astronomical Databases: GrayWulf-Connectivity



Almagest:
Potsdam, AIP

Disk-Server-Connection

Serial Attached SCSI (SAS)

- 1 to 3 Storageboxes (JBOD) per Server
- 1 Raid-Controller per Box

Server - Server Connection

Infiniband Socket Direct Protocol (SDP)

- 1 Inifiniband-HCA 20 GBit/s per Server

Internet - Infiniband

Ethernet to Infiniband via Router with

2x10 GBit/s Ethernet

2x 20 Gbit/s Inifiband

Hardware for astronomical Databases: GrayWulf-Architecture*

I/O from Hard-Disk to Raid controller

• harddisk has an I/O bandwidth of
75-80 MB/s (seq. read)
50-55 MB/s (write)

• for Raid 0 a JBOD with 15 disks should scale linear to 15*75 MB/s = 1125 MB/s (seq. read)

• one SAS lane has a bandwidth of 3 Gb/s, the connection to the raidcontroller on the host has 4 lanes: available bandwidth ~ 1200 MB/s

•Result: the JBOD/controller connection is sufficient to accommodate for the I/O bandwidth of ~1200MB/s

*A. Szalay et al., GrayWulf: Scalable Clustered Architecture for Data Intensive Computing

GrayWulf-Architecture*

I/O from Raid Controller to Infiniband

- current raid controllers saturate at 800-900 MB/s, introducing a serious bottleneck
- PCIx8 slots show 1600MB/s bandwidth
- The PCI bus saturates at 2400 MB/s
- Result: the PCI bus can accomodate for 2-3 RAID controllers
- Bandwidth for operations between
  CPU and Memory
  ~2400MB/s copy ,
  ~4100MB/s write,
  ~5700MB/s read
- Infiniband HCAs show 20Gb/s = 2500MB/s bandwidth

*A. Szalay et al., GrayWulf: Scalable Clustered Architecture for Data Intensive Computing

Hardware for astronomical Databases: GrayWulf-Architecture*

Basic ideas, starting from two of Amdahls‘ laws:

A balanced system needs

1 Bit I/O for each CPU cycle (IO)

1 Byte of Memory for each CPU cycle (BW)

The table below shows numbers for some current systems:

System	CPU count	GIPS [GHz]	RAM [GB]	diskIO [MB/s]	Amdahl	
					RAM	IO
BeoWulf	100	300	200	3000	0.67	0.080
Desktop	2	6	4	150	0.67	0.200
Cloud VM	1	3	4	30	1.33	0.080
SC1	212992	150000	18600	16900	0.12	0.001
SC2	2090	5000	8260	4700	1.65	0.008
GrayWulf	416	1107	1152	70000	1.04	0.506

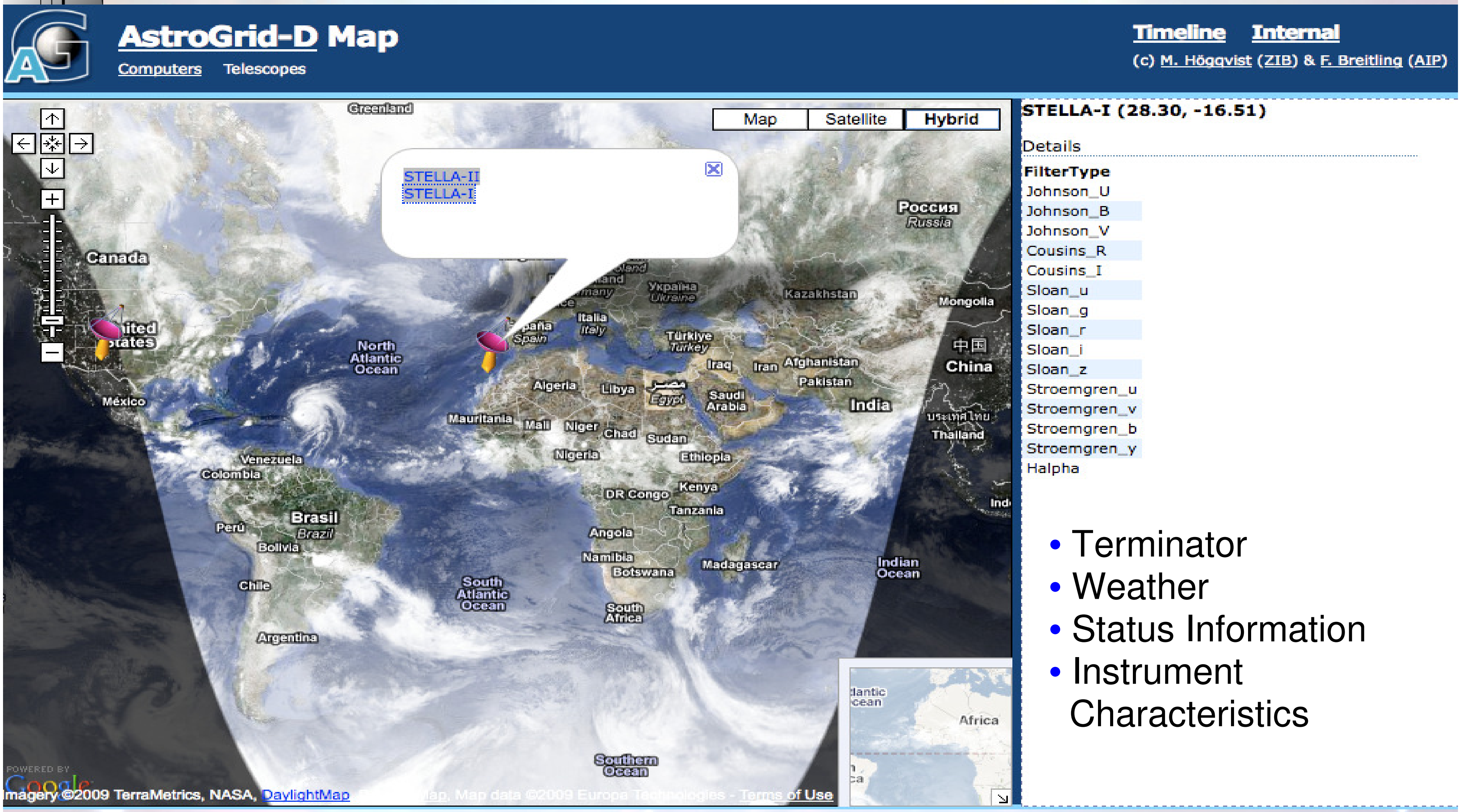
*A. Szalay et al., GrayWulf: Scalable Clustered Architecture for Data Intensive Computing

Robotic telescopes: Stella-I & Stella-II

- Teide Observatory, Tenerife, Canary Islands, Spain (2480m)
- Two 1.2m telescopes
- Instruments
  - Spectrograph
    (operating since May 2006)
  - Imaging photometer
    (commissioning in fall 2007)

- Scientific Objectives
  - Doppler imaging
  - Search for extrasolar planets
  - Spectroscopic surveys
  - Simultaneous observations with larger facilities

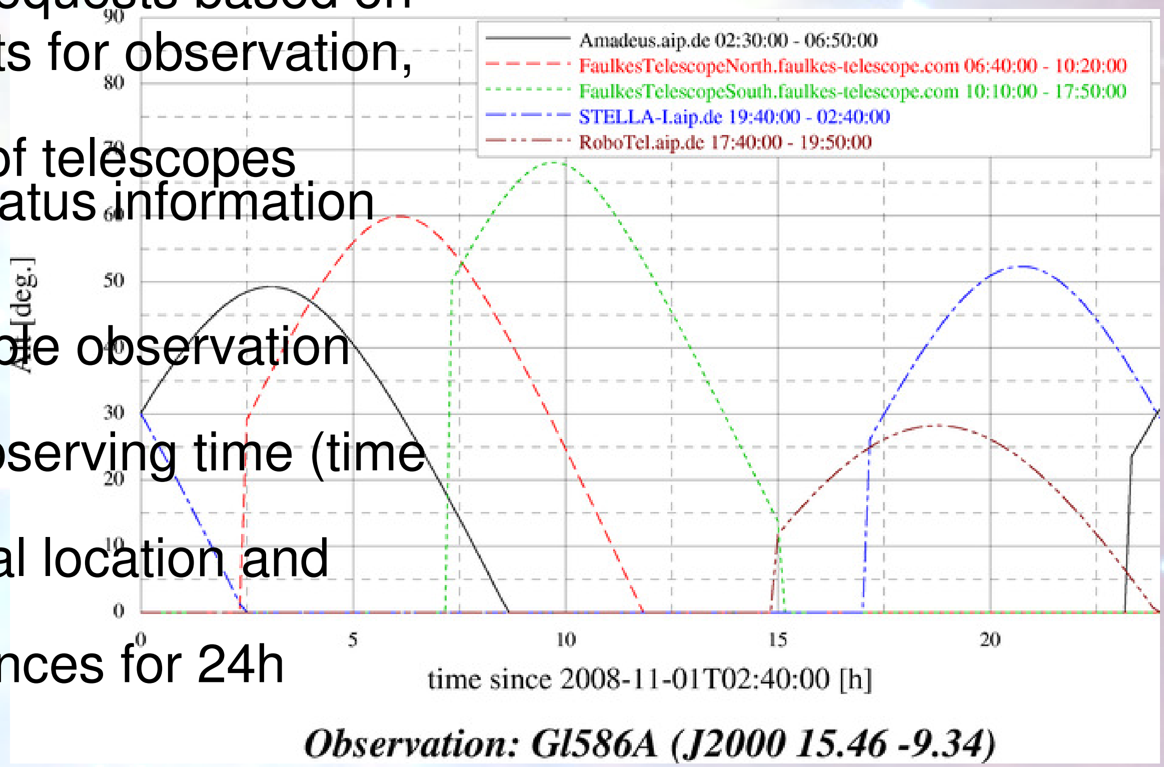
Robotic Telescopes as datasources



- Terminator
- Weather
- Status Information
- Instrument Characteristics

Components of the Robotic Telescope Network Package

- Broker
  - uses RTML profiles and templates for telescopes in the network
  - processes job requests based on
    - requirements for observation, e.g. filters
    - availability of telescopes based on status information
- Scheduler
  - produces possible observation schedules by
    - checking observing time (time of day)
    - geographical location and altitude
  - provides sequences for 24h observations



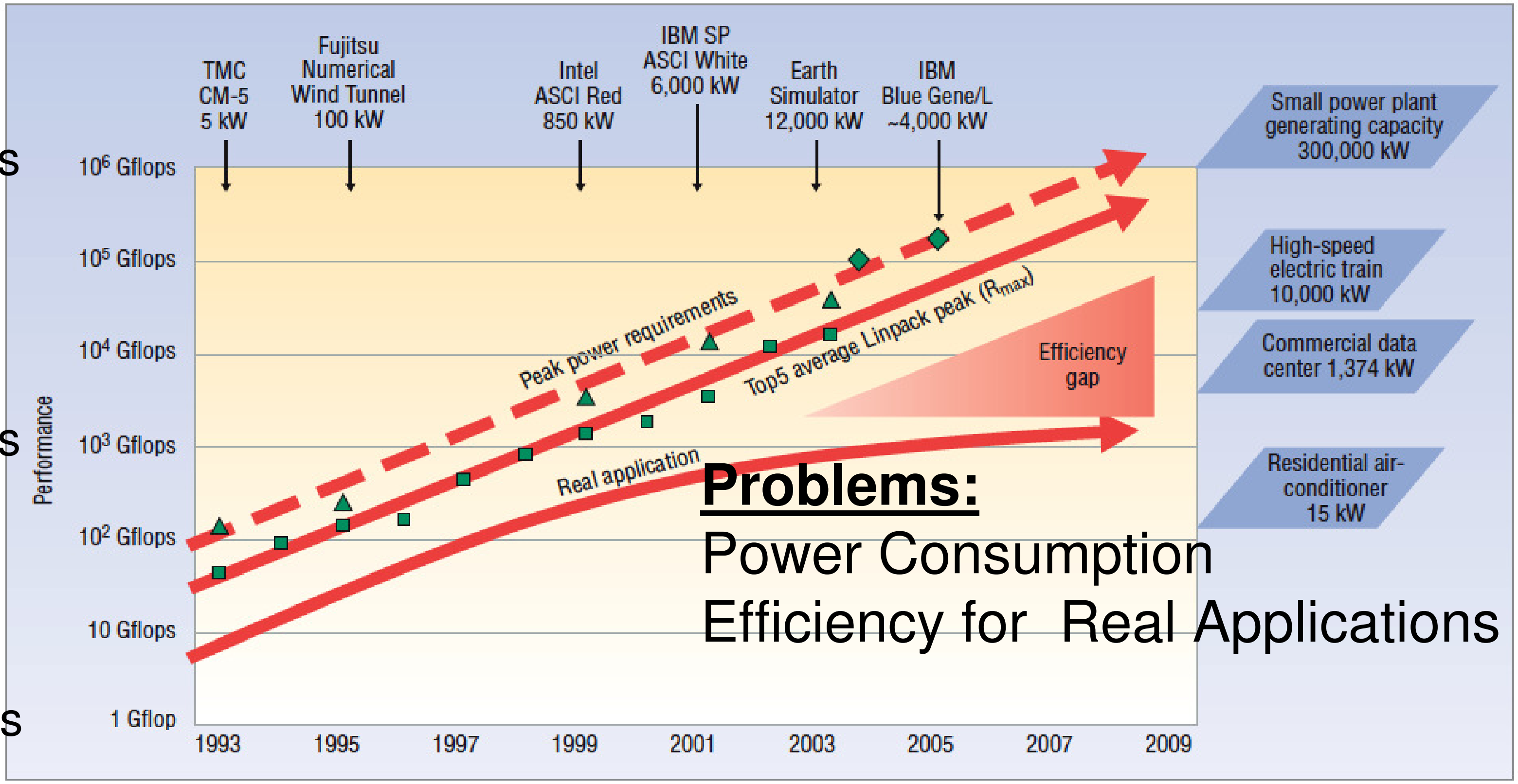
Green Grid / Green Supercomputing

Exaflop/s?

Petaflop/s

Teraflop/s

Gigaflop/s



Problems:

Power Consumption

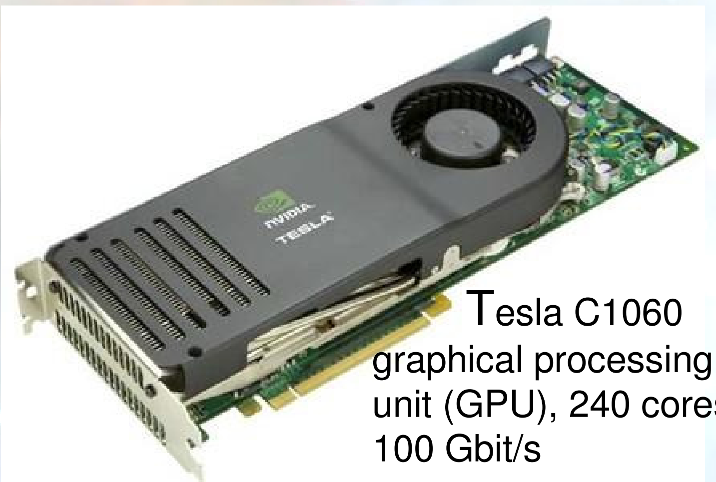
Efficiency for Real Applications

Figure 1. Rising power requirements. Peak power consumption of the top supercomputers has steadily increased over the past 15 years.

Thanks to Horst Simon, LBNL/NERSC for this diagram.

Energy efficiency

Beyond GRAPE solutions....



Tesla C1060 graphical processing unit (GPU), 240 cores, 100 Gbit/s

<u>GPU 1 Tflop/s peak ~ 0.4 Watt / Gflop/s</u>

IBM BlueGene ~ 2.5 Watt / Gflop/s

Standard PC ~ 25 Watt / Gflop/s

<u>Real Codes Need More Power:</u>

Efficiency N-Body ~ 80 % ~ 0.32 W/Gflop/s

Efficiency SPH ~ 4 % ~ 10 W/Gflop/s

<u>Reconfigurable Hardware (FPGA)</u>

Univ. Heidelberg MPRACE-1 Board

Efficiency SPH ~ 4 W/Gflop/s

Courtesy of R. Spurzem

Green Supercomputing

... is application driven:

- Cosmological Structure Formation and Galaxy Formation *(NAOC: Gao Liang)*
- Star Cluster Dynamics *(KIAA, Stephen Justham, Sverre Aarseth,...)*
- Binary Black Holes, Galactic Centres, Gravitational Waves
  (NAOC, KIAA/PKU, UHD: Rainer Spurzem, Peter Berczik, Fukun Liu...)
- Planet Formation and Evolution
  (KIAA/PKU: Doug Lin, Rainer Spurzem, ...)
- Dome A Project Antarctica Observational Data *(NAOC: Lifan Wang, ...)*
- Molecular Dynamics / Chemical Processes *(IPE/CAS)*
- Numerical Relativity? *(Inst. Appl. Math. CAS)*

Green Grid of Special Supercomputers

Beijing (NAOC/IPE/CAS)
Heidelberg (ZAH, ZITI)
Berkeley (LBNL, C3)

1) Support Key Applications / Computational Challenges at NAOC/KIAA/PKU with Green Supercomputer
2) Test early application driven use of future technologies in Green Supercomputing, support early production systems
3) National and International Cooperation to support common high level Software Interfaces, Grid Interfaces.
4) Explore new technologies (FPGA Computing, HT network)

5) Collect Benchmarks and key applications on three continents

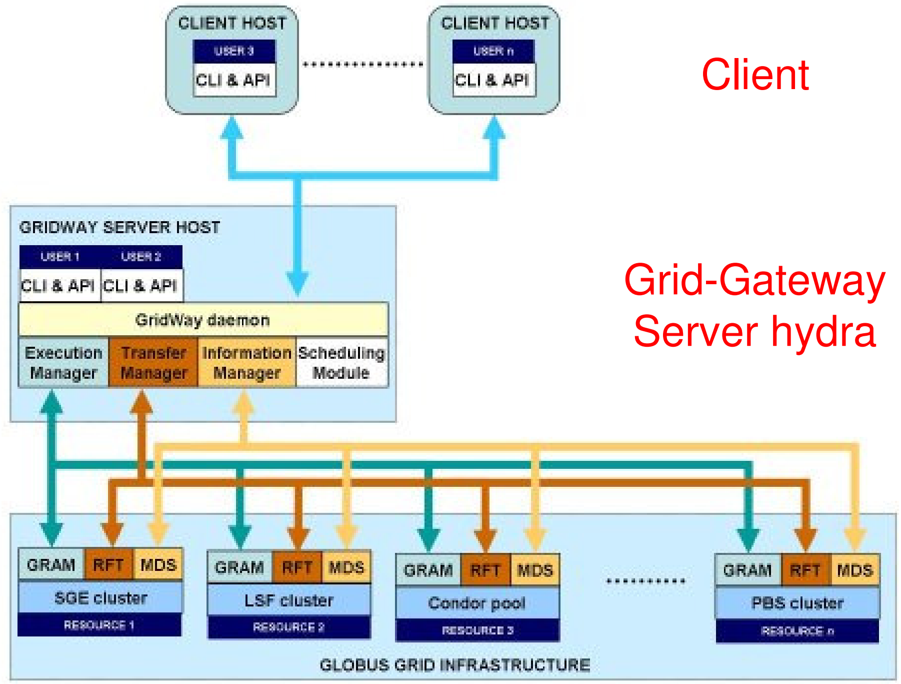
Building on experiences of Astrogrid-D/MAO-NASU :

- Globus middleware installation
- VO-Management w. VOMRS
- forming Virtual Organisation as base for easy collaboration
- domain: www.gpu-grid.org

AstroGrid-D

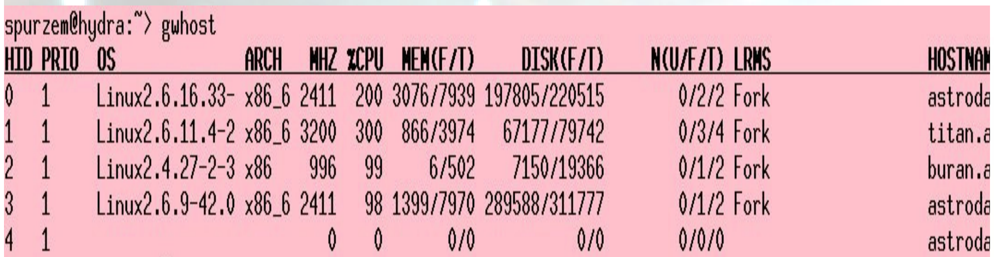
The Astrogrid-D currently offers: software to access the grid:

- Web Portal to access the grid for common applications
- Gridway / Grid-Gateway for classical job submission



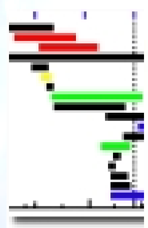
Client

Grid-Gateway
Server hydra



```
spurzem@hydra:~> gwhost
HID PRIO  OS              ARCH   MHZ %CPU  MEM(F/T)    DISK(F/T)     N(U/F/T) LRMS       HOSTNAM
0   1     Linux2.6.16.33- x86_6 2411  200 3076/7939 197805/220515     0/2/2 Fork       astroda
1   1     Linux2.6.11.4-2 x86_6 3200  300  866/3974   67177/79742     0/3/4 Fork       titan.a
2   1     Linux2.4.27-2-3 x86    996   99     6/502    7150/19366     0/1/2 Fork       buran.a
3   1     Linux2.6.9-42.0 x86_6 2411   98 1399/7970 289588/311777     0/1/2 Fork       astroda
4   1                              0    0       0/0           0/0     0/0/0            astroda
```

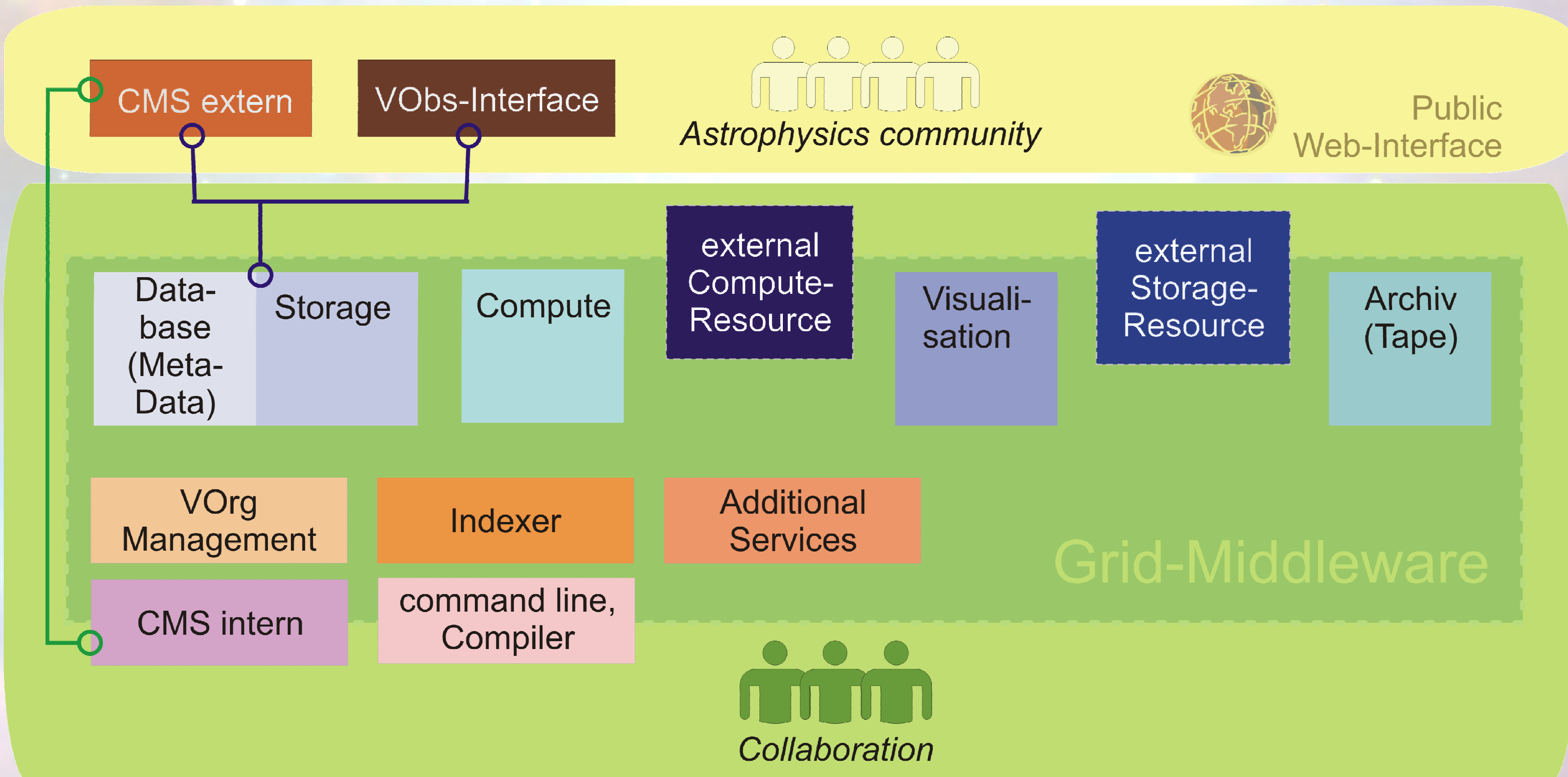
Stellaris



Information Systems | Gridmap | Grid Timeline | Astrogrid-D Resources

VRE: Collaborative Research Environments and Networks

Elements of a Collaborative Research Environment
(Example: Simulation Postprocessing)



Virtual Research Environment

Dedicated for use by collaboration
-machine (16 Cores (2.4 Intel), 128GB Memory)
-planned for up to 100TB storage
-SAS RAIDs, Infiniband connect

Content Management System for project with
- public site + intranet site
Additional webservers, mailing list
ssh, grid + hpn-ssh interfaces available
compiler, idl

under development: database for
registering simulations + postprocessing work
planned:
other support tools (e.g. metadata generation, visualisation)

VDZ: Virtual Research Environment

GPU - GreenGrid - VRE

- **Specialized Hardware**
  - **Provides powerful number crunching capabilities at (almost) commodity prices**
  - **Reduces power consumption**
  - **Enables new modi of use**
  - **Provides required capabilities for astronomical data archives and data mining tools**
  - **Opens up new areas of collaboration**
- **Collaborative use of these resources**
  - **by (Green) Grid and other available Grid Infrastructure**
  - **by optimized Virtual Research Environments**

is one of the future key elements of scientific work

Grid Integration