I/O@HLRS

Thomas Bönisch

I/O @ a Hotel

- Everybody talks about capacity
  - Gigabyte, …
- Throughput problems
  - I/O comes in bursts
  - I/O queues

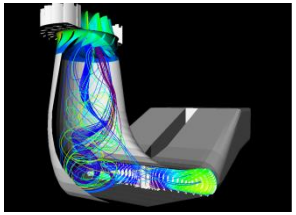
Outline

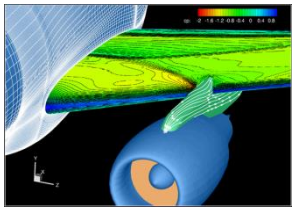
- HLRS
- Applications & Usage
- Hornet (the system)
- I/O Architecture
- Monitoring
- Issues and solved problems
- HSM
- Future Perspectives

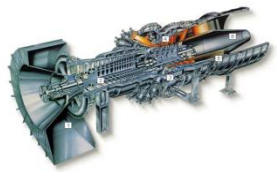
The High Performance Computing Center Stuttgart (HLRS)

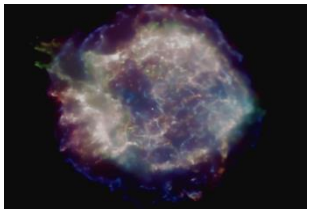
- Central Unit of Universität Stuttgart
  - Supercomputing since 1968
- 1st German National Supercomputing Center
  - Founded 1996
  - service for German researchers
- Gauss Center for Supercomputing
  - Founded 2007, Partners: Jülich and Munich
- Open for European users since 2004
- Partner for German industry

Main Areas of Users' Research

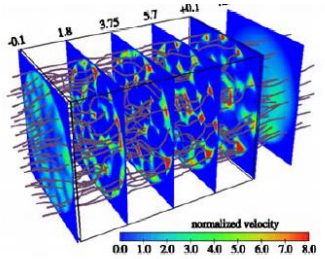


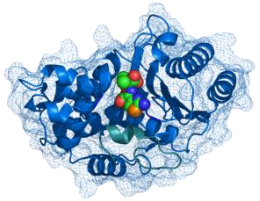


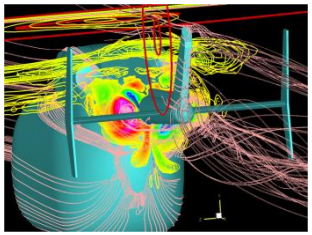




- Aeroacoustics
- Aerodynamics
- Astrophysics
- Bioinformatics
- Combustion
- Fluid-Structure Interaction
- Helicopter Aerodynamics
- Meterology
- Medical Imaging
- Nanotechnology
- Solid State Physics
- Turbo Machinery
- Turbulence Phenomena



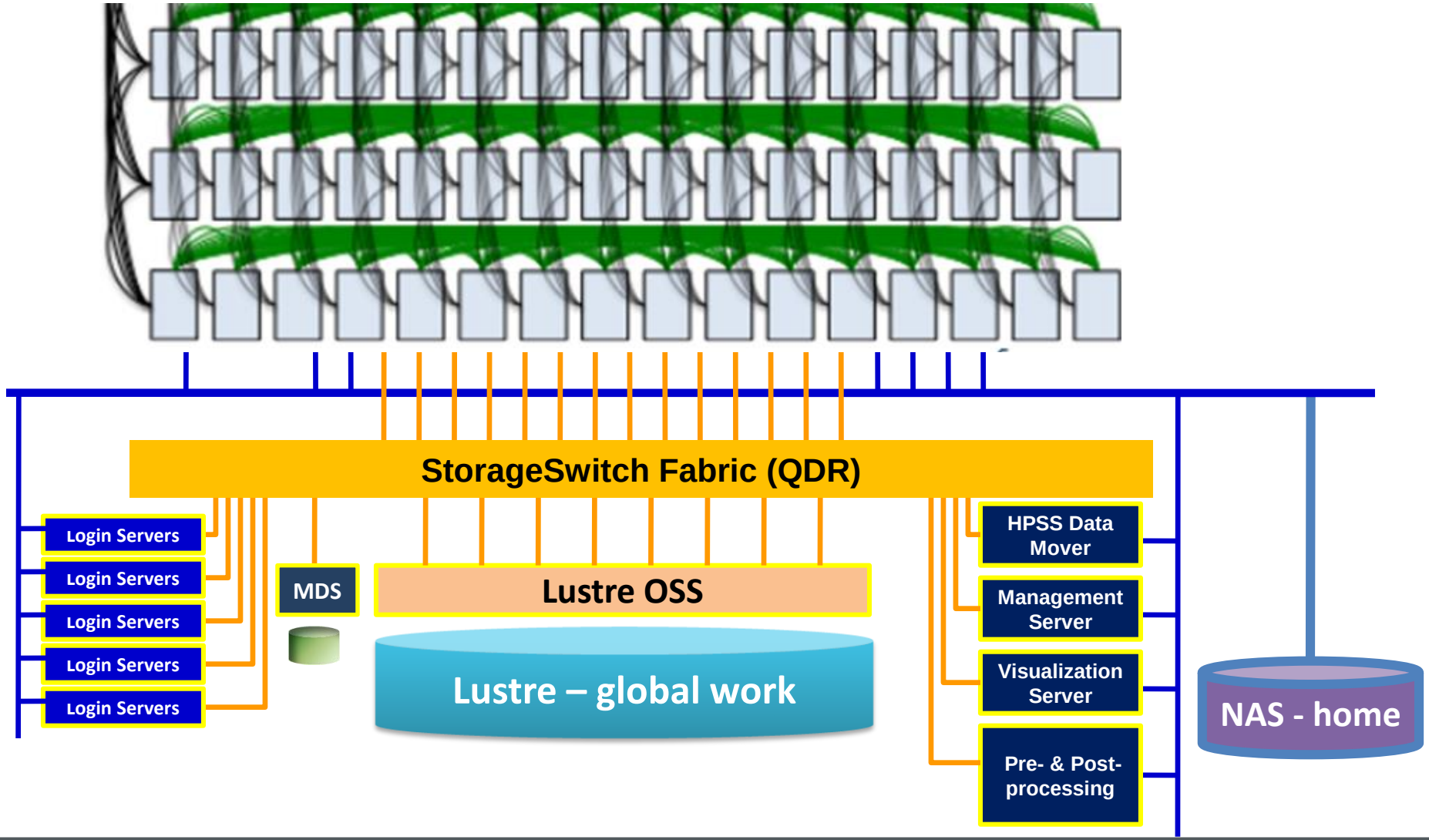




HLRS Cray XC40 (Hornet): Overview

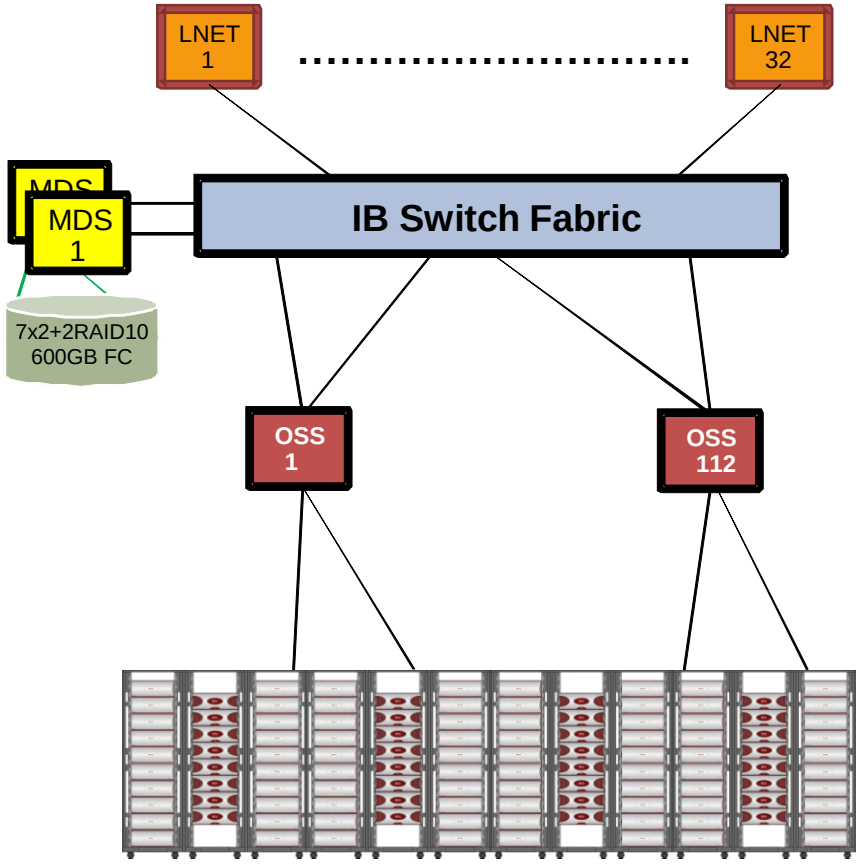
- Configuration:
  - Peak Performace ~3.786 Petaflops
  - 3944 compute nodes
  - 2 sockets / node
    - Intel Xeon E5-2680v3 (Haswell@ 2.5GHz 12 Cores each) leading to 94,556 cores
  - 128 GB main memory per node (5.3 GB/core) → 493 TB in total
  - Aries network
  - ~2MW maximal power consumption

Conceptual Architecture



I/O architecture

- Hardware
  - 7+7 MDS/MGS Servers
  - 112 OSS Servers
  - 22 Dual RAID controllers
  - 7440 Hard disks
- 2.7 PB Storage of 2011
  - 120 GB/s measured
- 5.4 PB Storage of 2014
  - > 150 GB/s measured



3840TB + 7200 TB Total Raw Capacity

Work Space Mechanism

- A directory in the project file system is created upon request with a user defined name
- The directory is available for 30 days
- The directory life time can be extended 3 times by 30 days
- At the end of life, the directory with its content!!! is automatically deleted
- There are tools for
  - finding available workspaces
  - Releasing workspaces
  - Setting a reminder in calender tools
- Quota is enabled

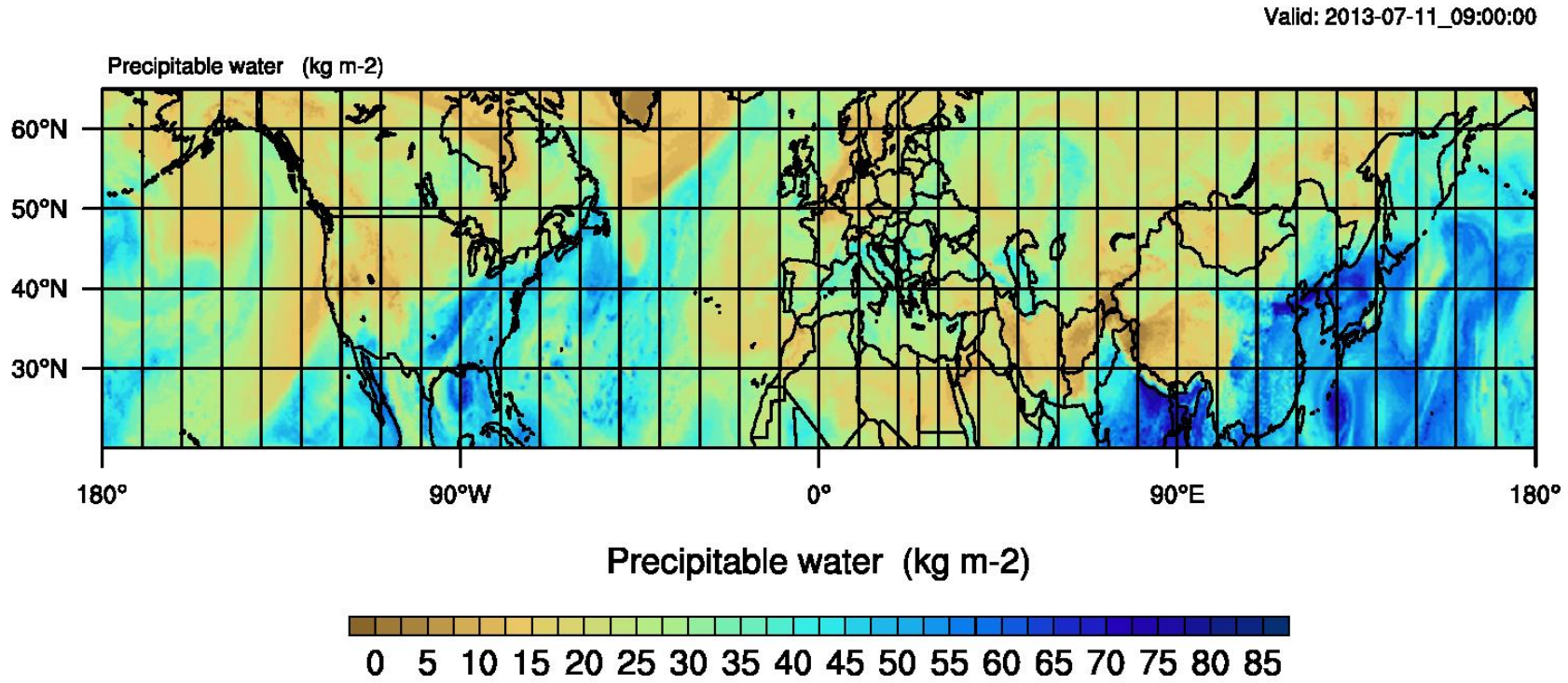
Usage numbers & issues

- One file system for general usage
  - 3.5 mio files
  - ~ 500 TB usage (out of ~700 TB)
- Other file systems by invitation only
  - Power users (capacity, throughput)
  - Industry
- Issues
  - Small files
  - I/O performance of application is rarely looked into

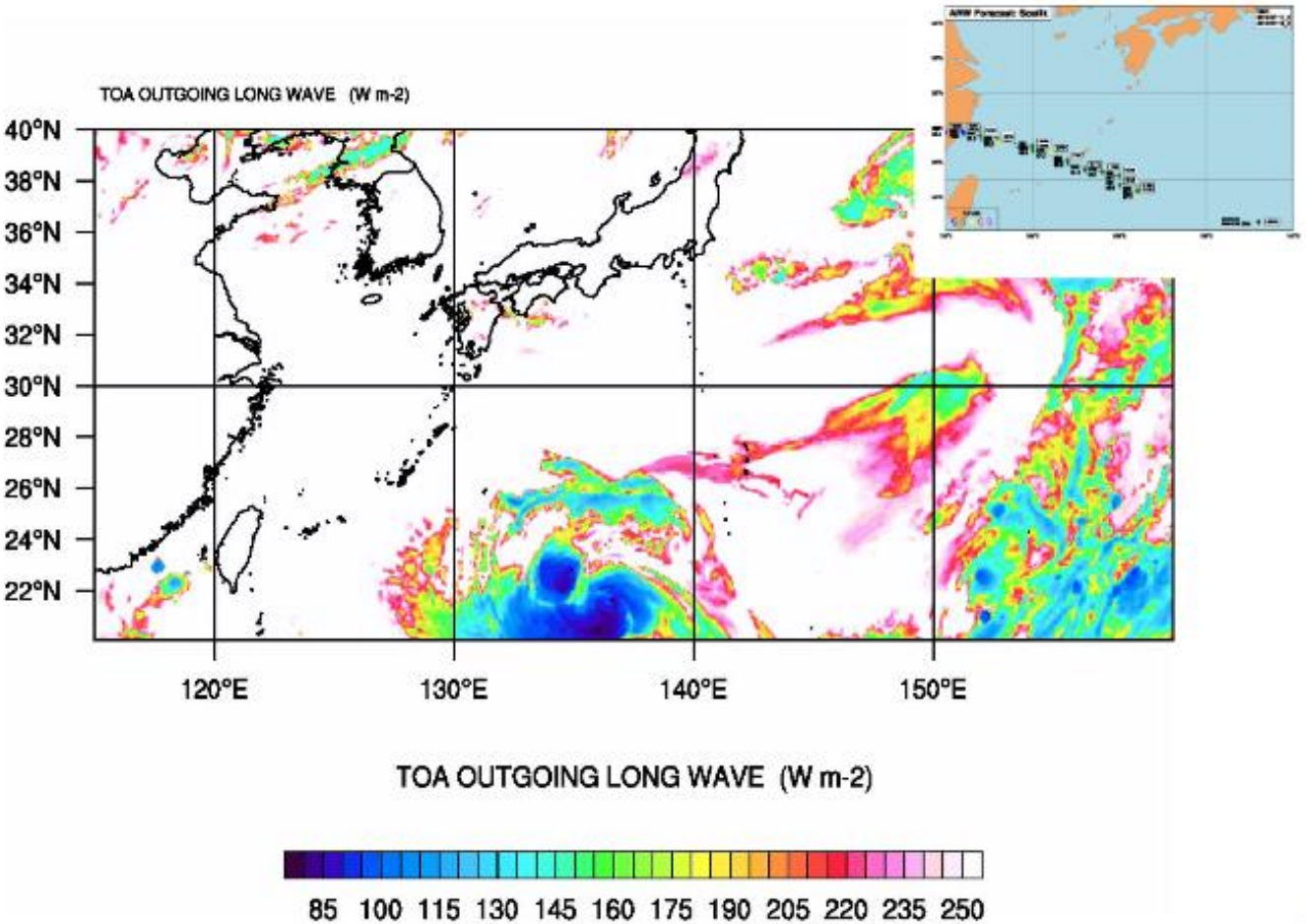
Convection permitting Channel Simulation

- Institut für Physik und Meteorologie, Universität Hohenheim
  - Wulfmeyer, Warrach-Sagi, Schwitalla
- WRF model, 3.3 km resolution
- 3500 nodes=84000 cores; 330 TB data; 84 system hours
- Vertically integrated water vapor nicely shows the fine scale structure of the atmosphere.
- Visible is the Monsoon circulation over India, Typhoon Soulik close to Taiwan and a tropical depression in the Gulf of Mexico.
- The sharp gradient of moist air masses over the North Atlantic is also visible. Low pressure systems influencing Europe are developing along this line.

Convection permitting Channel Simulation

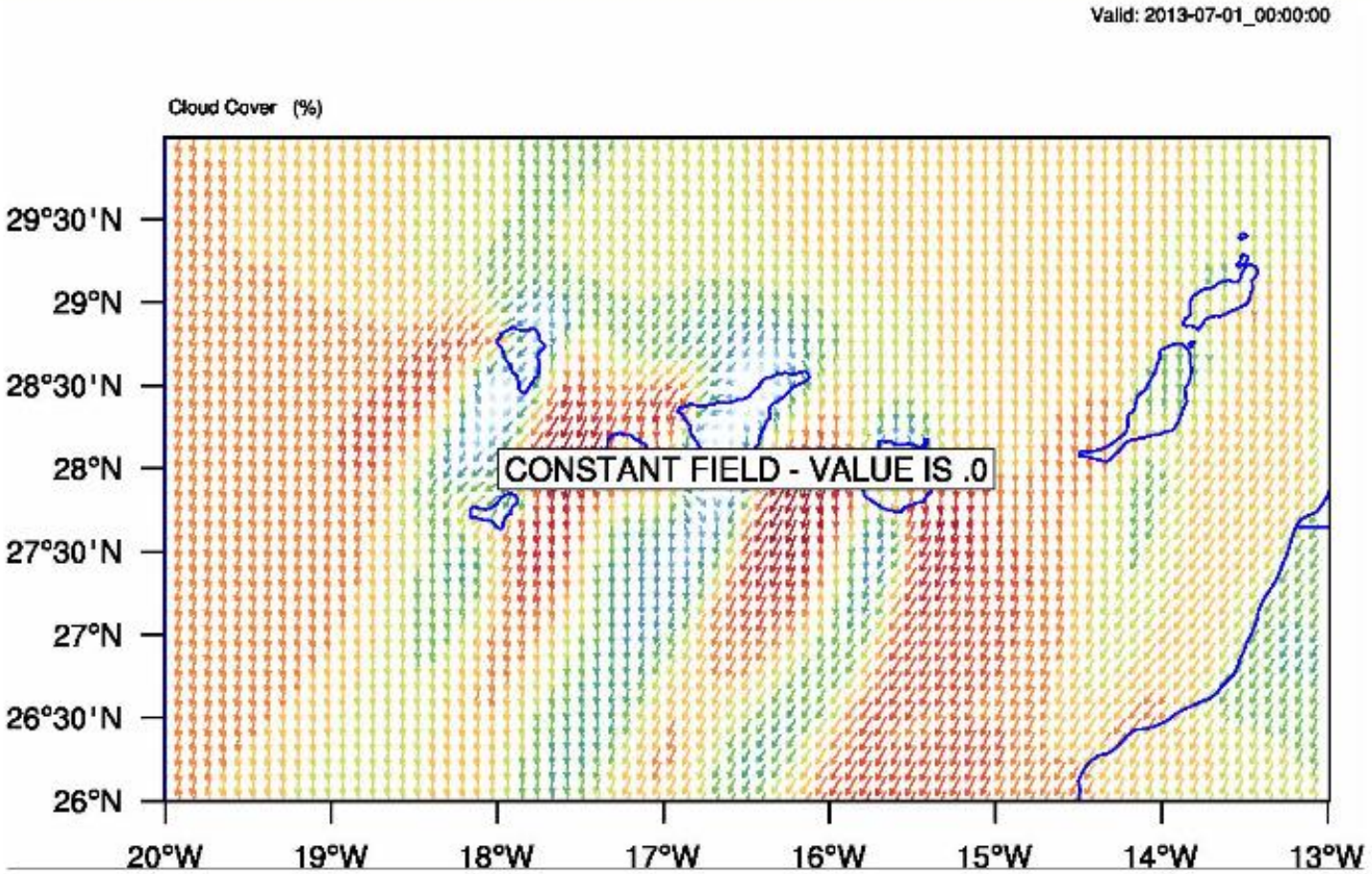


Typhoon Soulik



UNIVERSITY OF HOHENHEIM

Karman Vortex Street



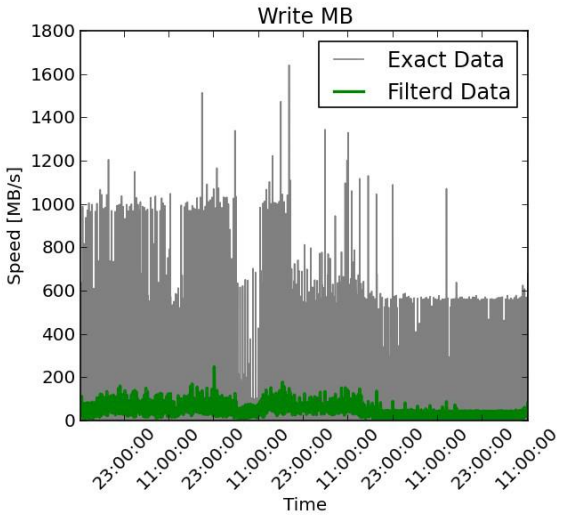
The issues

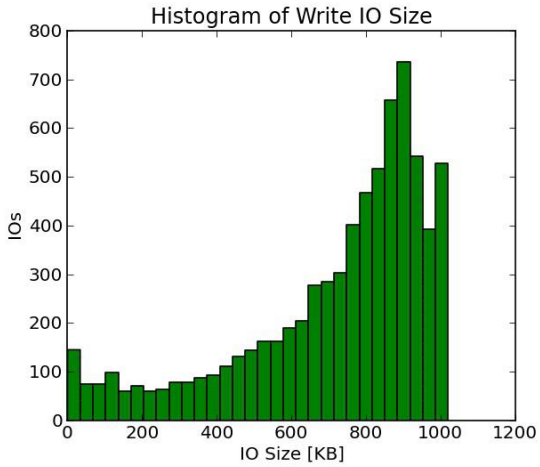
- I/O was and is a problem of this code
- 1st shot:
  - 1 GB/s throughput
- After optimization
  - 7.5 GB/s throughput
  - 2 days of calculation.
  - 1.5 days of I/O
- File System potential: 75 GB/s (measured !!!)
- Software: netcdf4

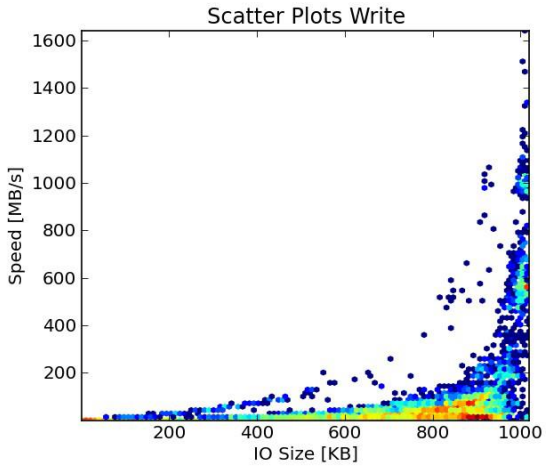
Monitoring

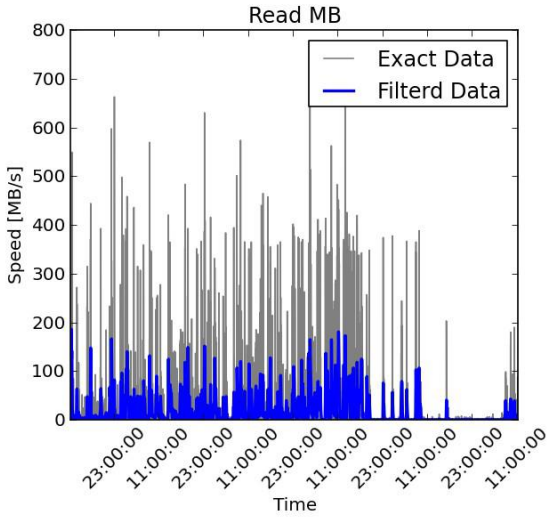
- Self written tool
- Reads lustre performance counter
- Stores data at an external server in a data base
- Combines information with accounting data
  - User information
  - Job information: Run time, start time, end time, etc.

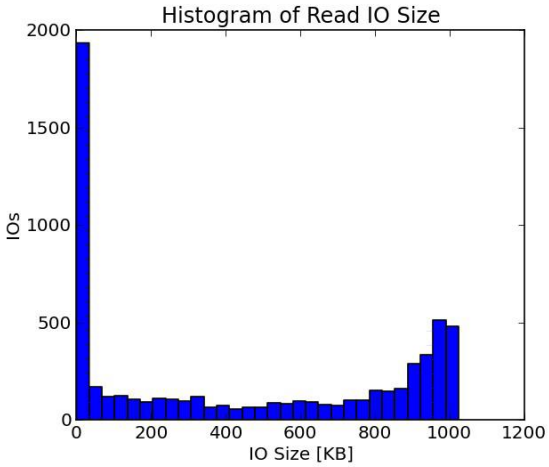
Performance data

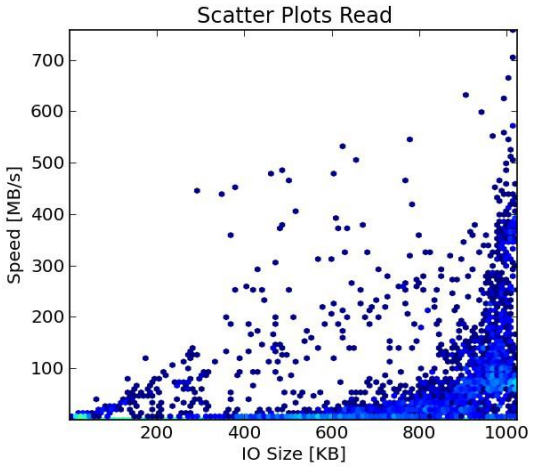












Issues

- Hardware
  - Data loss because of a system miss configuration (active – active controllers)
  - Unexpected influences between file systems.
  - 15 days operational issues
- Software (Lustre)
  - Quota
  - Upgrade procedures
  - Short read
    - Netcdf, Fortran, pftp_client
  - OSTs being close to full have interesting effects when being used with ISV codes
- Software (Libraries)
  - Performance of libraries: Netcdf, …

Outlook

- Extension of the Cray XC40
  - Hazel hen; Q3/2015
  - Peak Performace ~7,4 Petaflops; 185.376 cores
  - 965 TB Main Memory
- Additional Storage
  - Cray Sonexion with 13 SSUs
  - 97,5 GB/s; 5 PB usable storage

The Future

- How will NVRAM change the game?
  - Architecture
  - Interfaces
    - Block (for compatibility)
    - Byte addressable (malloc)
- My vision
- What does this mean for users?