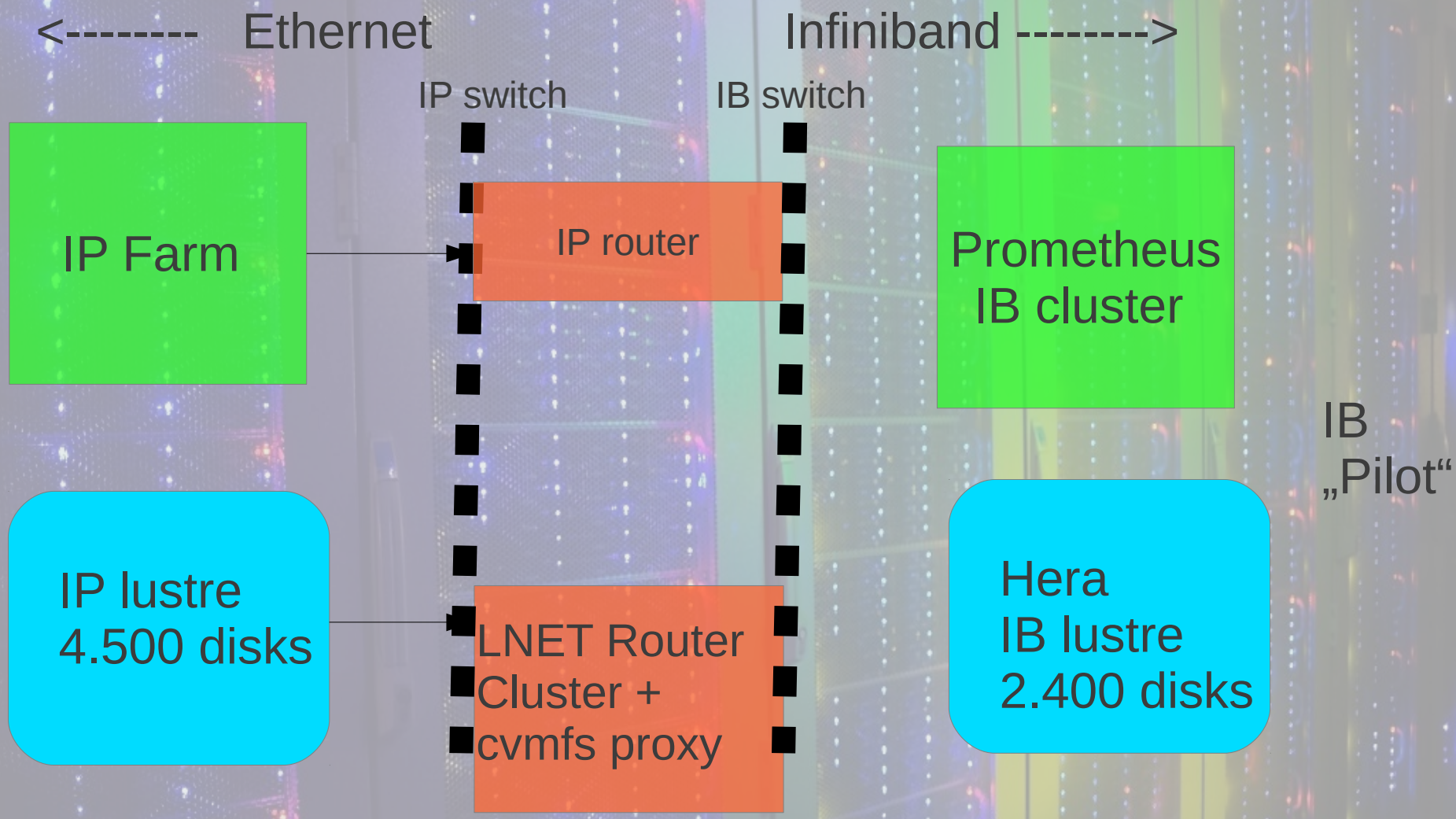
Lustre Cluster
Hera
Prague - spring 2012

Walter Schoen, GSI

Hera – the IB lustre part

- IB based, commodity hardware (cheap :-))
  Even boot is over IB (low latency booting :-))
- Installation & configuration in a couple of days
  thanks to our Chef based fabric management :-)
- Independent Ethernet network for IPMI Modules
- Connected via LNET router cluster to IP lustre
- „Pilot“ Version with 2.400 disks
- Two independent „thermal measures“
  - System temperature based on IPMI Modul
    => graceful shutdown if threshold reached
  - Rack temperature inside/outside door above thresh.
    => emergency shutdown of PDU



Lustre at GSI: about 7000 disks, 6.5 PB Raw
150 Fileserver, about 1 Tb/s I/O

Prometheus / Hera
- IB cluster in the Minicube

Executive summary: works! Success! :-)

- **Prometheus: 10,000 core IB cluster pilot**
- **Hera:IB pilot with 2,400 disks lustre, 50 File servers, will be extended soon ….**

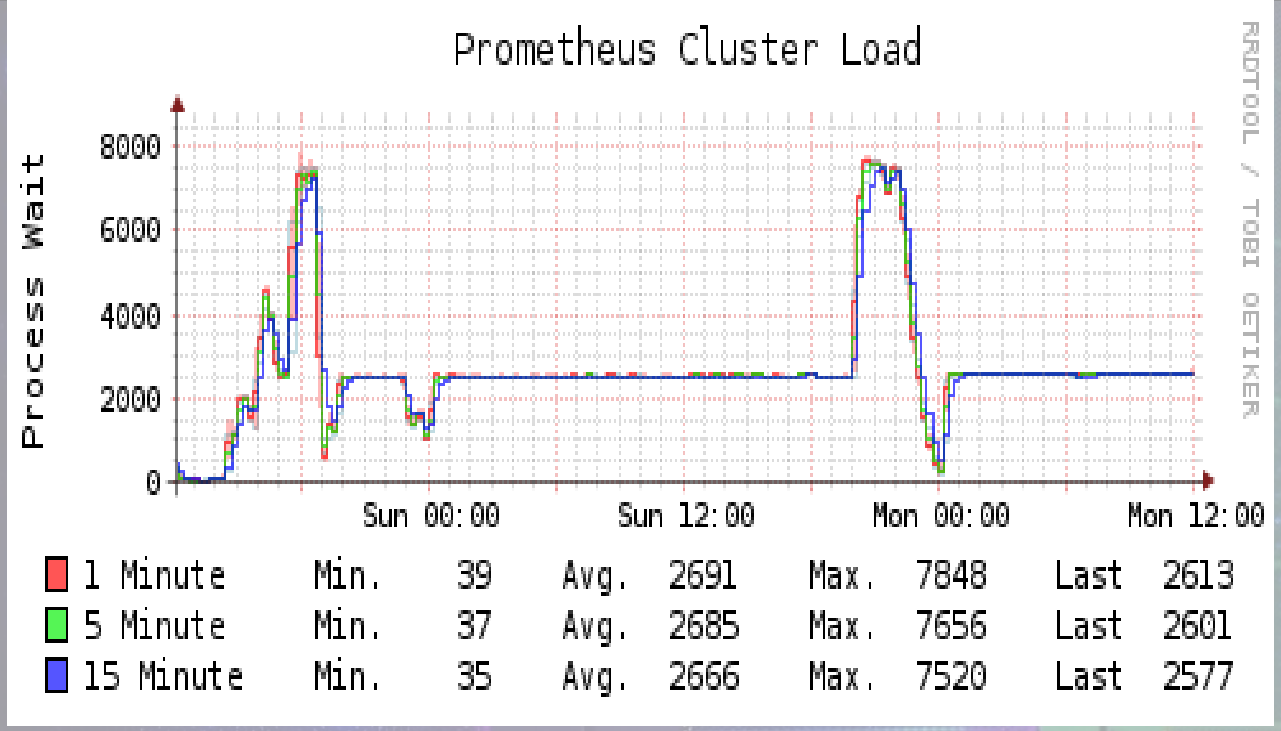
Prometheus + Hera – first results

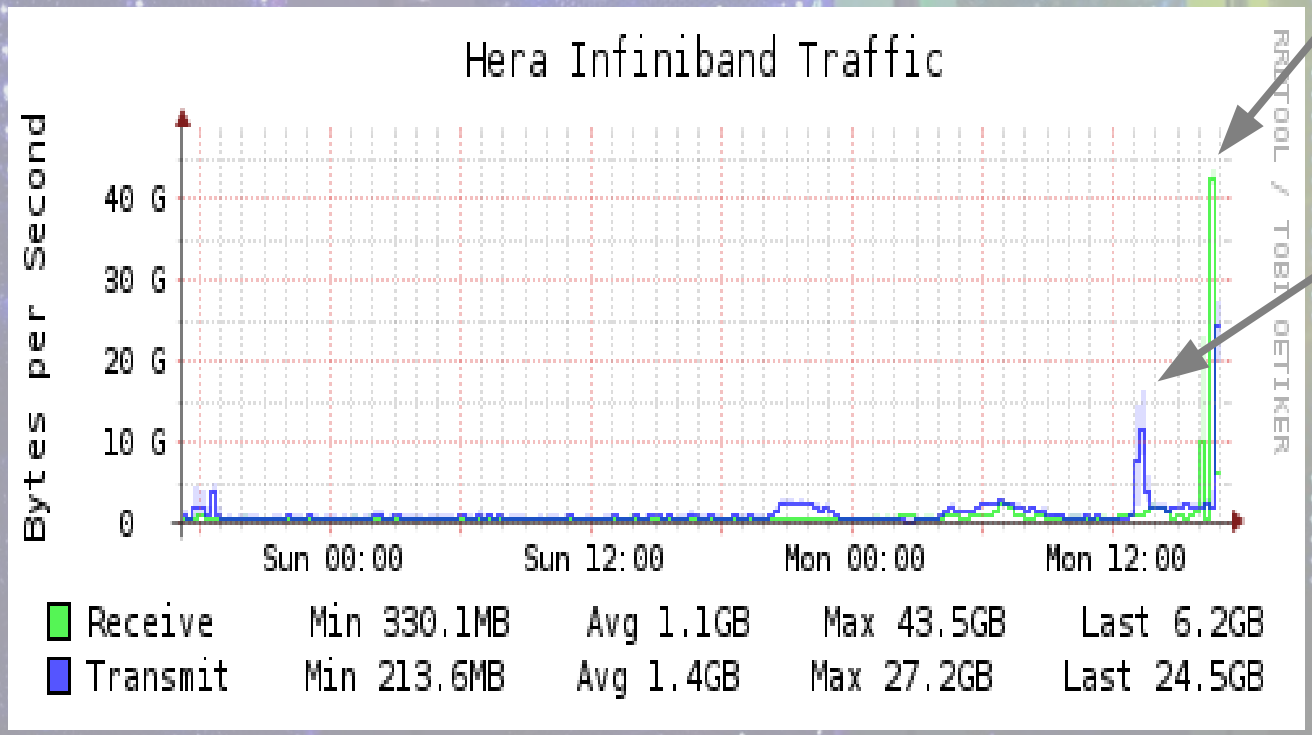
- Few weeks in operation now
- 1.000.000 physics jobs already finished
- 6 PB data throughput per week (few users ...)
- ~.5 Tbit/s I/O troughput already reached
  (in the Hera part, limited by the switched backplane)

=> **Already (successfully) used by**
Alice, Acclerator-Physics, GSI-Theory, HIM

Prometheus + Hera – first results (out of the box :-))

- IB works reliable
- Speed measured with Prometheus
  - RDMA per node: 3.1 Gbyte/s (expected: Mellanox: ~3.1GB)
  - Parallel access to lustre per node: Read ~2 Gbyte/s
    about a factor 20 faster compared to our IP clients.....
- Fileserver performance now limited by the switched backplane..





HPC Test
450 Gbit/s
(close to the backplane limit)

Alice Train,
160 Gbit/s

File Edit View History Bookmarks Tools Help

http://lxrm02.gsi.de:5000/report?display=gridengine%2Faccounting

Most Visited Getting Started Latest Headlines

Report

Graph | Data | Report | Help

Select reports to display:

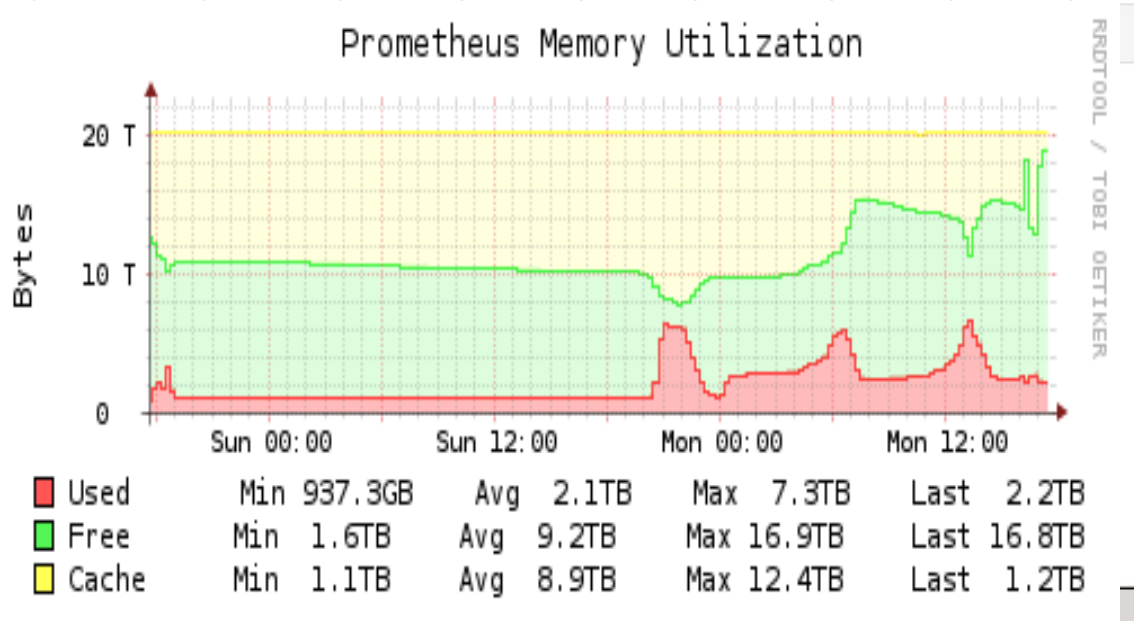
all

Show Cancel

User	Jobs	CPU Time	Run Time	Memory Used Over Time	I/O	Jobs Max Resident Memory <1GB	<2GB	<4GB	>4GB	Jobs Run Time <1h	<4h	<12h	>12h
wiechula	11563	14kh	15kh	65.53PB/s	190.62TB	7521	3747	284	11	6317	4185	1061	0
tneff	6	1h42m	5m37s	1.22TB/s	31.88MB	6	0	0	0	6	0	0	0
vpenso	9	0s	8s	0.0B/s	6.27MB	9	0	0	0	9	0	0	0
root	2	0s	2s	0.0B/s	1.21MB	2	0	0	0	2	0	0	0
sma	192949	37kh	115kh	80.08EB/s	2.03PB	140600	51555	193	601	178022	14920	7	0
jthaeder	104105	233kh	242kh	17.27EB/s	3.73PB	80305	0	23800	0	53333	26902	23869	1
mfasel	20	3h58m	4h11m	27.88TB/s									

/report/gridengine/accounting?format=html

Few days,few users:
6 PB I/O .....



Done


Next Steps:
add more discs soon …..