Disk to Network Streaming at 40 Gbit/s

Ronald van der Pol

<rvdp@sara.nl>

GLIF Workshop, 13-14 October 2010, Geneva, Switzerland

Outline

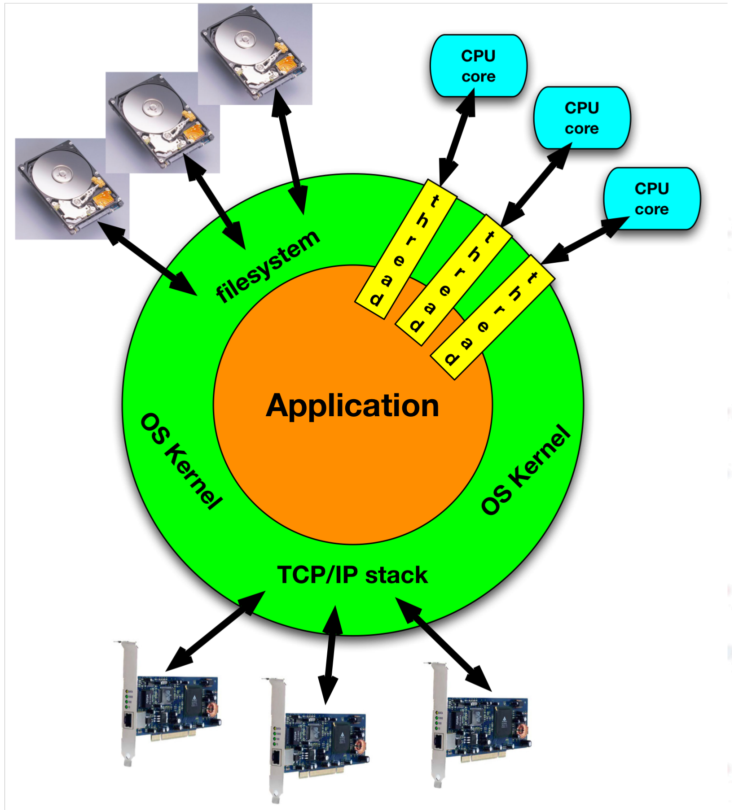
- Goal of this project
- 40G demonstration setup
- Application description
- Results
- Conclusions

Goal of the project

- Optimize single server disk to network I/O
- Optimize throughput from application to 40/100 Gbit/s transport networks
- Use mainstream hardware, no complex grid clusters
- Make use of parallelism (multiple disks, multiple cores, multiple NICs)
- Understand server architecture and compose a balanced server
- Make sure that disk I/O matches network I/O
- Avoid CPU bottleneck (enough cores)
- Avoid internal bus bottlenecks
  - Between memory and CPU
  - Between disk and CPU
  - Between NIC and CPU

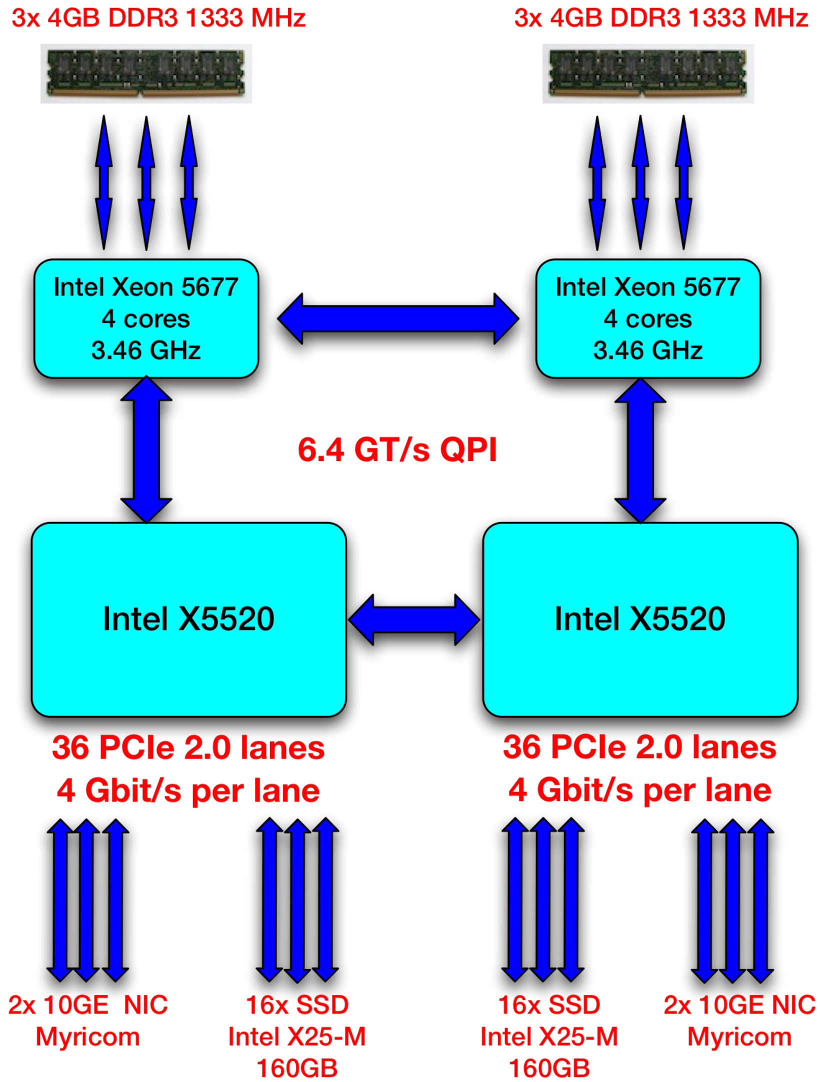
I/O Scalability

- Storage I/O speedup with multiple disks (RAID-1/RAID-Z)
- Compute speedup with multi-core systems
- Network I/O speedup with multiple NICs

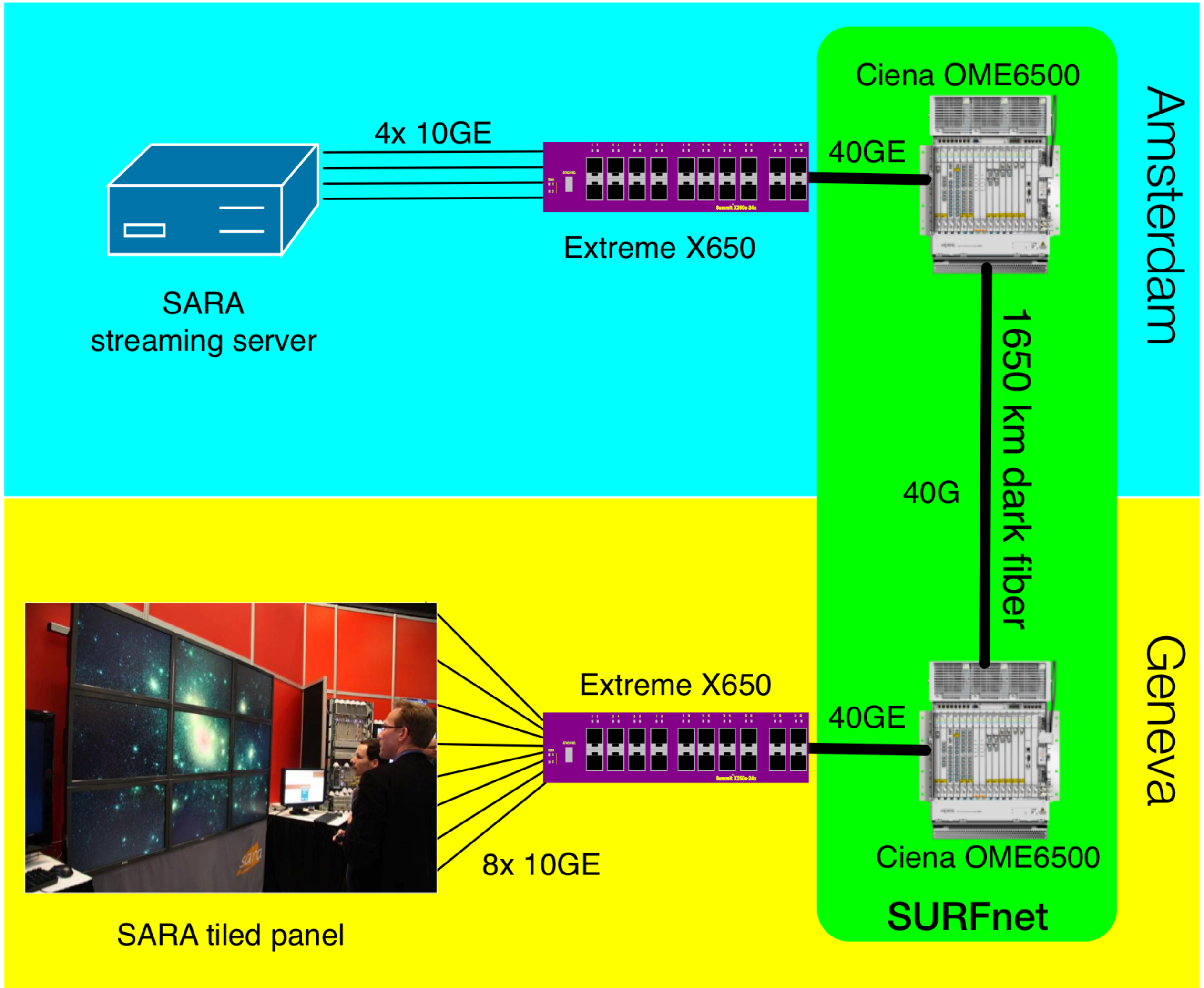


Streaming Server Architecture

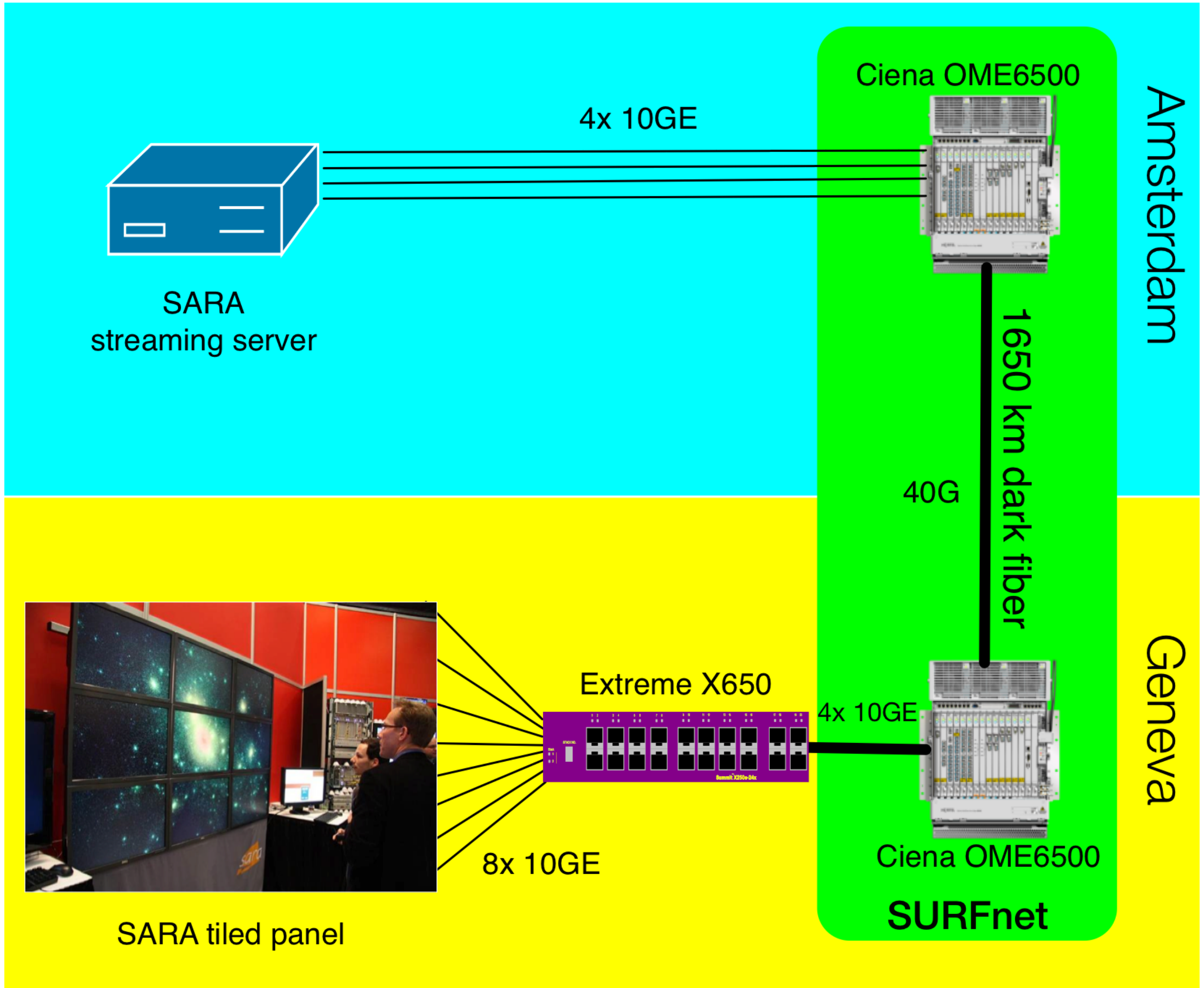
Supermicro X8DAH+-F motherboard
2x PCI-E 2.0 x16
4x PCI-E 2.0 x8
1x PCI-E 2.0 x4
8 cores @ 3.46GHz (Intel Xeon 5677)
24GB DDR3 @ 1333 MHz
4x 10GE Myricom (dual port)
32x SSD Intel X25-M 160GB



Planned 40G Demo Topology



Actual 40G Demo Setup



Streaming Application

- Streaming of single server to 5x3 Tiled Panel Display (TPD)
- 5x3 TPD has 15 LCDs
- Application runs on 1 streaming server in Amsterdam
  - Application spawns 15 MPI threads
  - Each thread reads data from disk and streams to an LCD
- CosmoGrid movies stored on SSD disks as 24bit RGB
- Streaming is done with UDP
- UDP streams are balanced over 4 NICs in streaming server

Tiled Panel Setup

- 5x3 TPD with 2560x1600 pixel LCD screens
- Total of 12,800 x 4,800 pixels (61.44 Mpixels)
- 8 servers
  - 7 with 2 LCD screens
  - 1 with 1 LCD screen
- 15 UDP streams
- Each server has 1x Nvidia GeForce GTX460 video card
- 1 TCP control stream for joystick

CosmoGrid

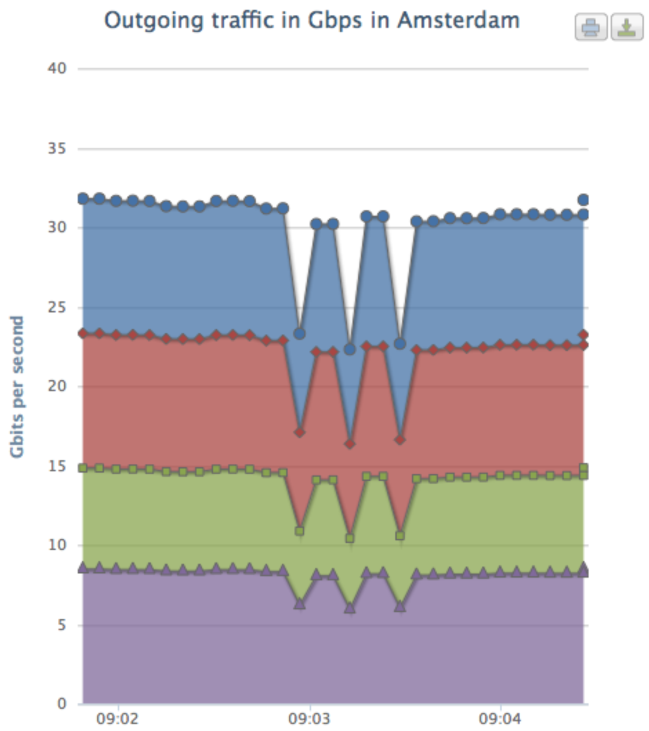
- Dutch computing challenge project (prof. Portegies Zwart)
- Simulation of 256^3 and 2048^3 bodies of dark matter
- Simulations shows formation of clusters after big bang
- Distributed application using several European and Japanese supercomputers

Results

- 32 Gbit/s disk to network from a single server
- SSD read speed with 16 disks: 2750 MiBytes/s
  - 172 MiBytes/s per disk
  - 256 MiBytes/s on a single disk
- Streaming consumes 142 Watt on streaming server
- CPU ~ 70% busy

Network Performance

SARA - LIVE 40GE monitoring

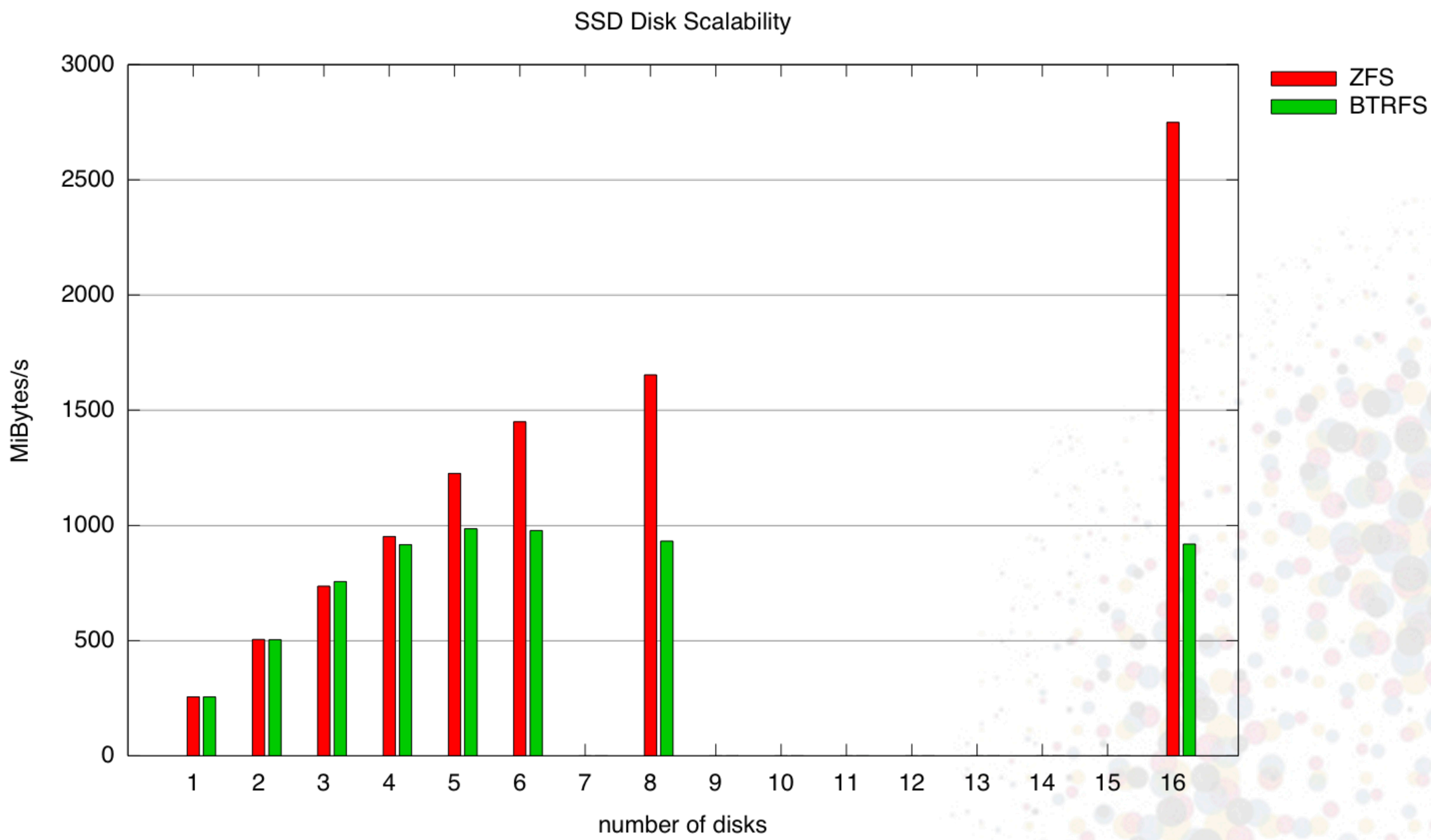


Supported by

OS Treadeoffs

- ZFS is not implemented in Linux
- ZFS is implemented on (Open)Solaris and FreeBSD
- BTRFS is supposed to be the Linux equivalent of ZFS
  - But ZFS still scales much better than BTRFS
- We had trouble getting SAGE running on (Open)Solaris and FreeBSD
- So we ended up with Linux and XFS

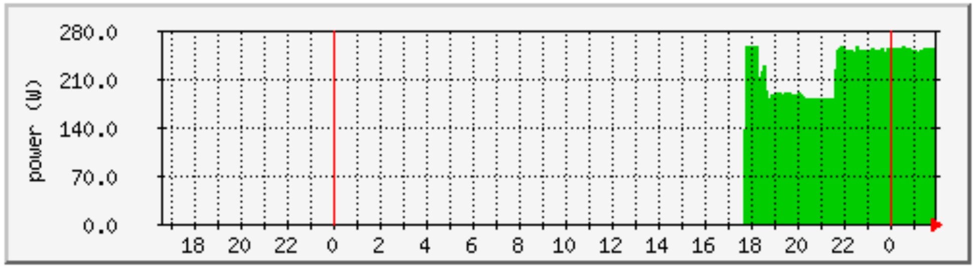
SSD Disk Scalability



Power Usage Streaming Server

The statistics were last updated **Friday, 1 October 2010 at 1:58**, at which time **'asd-powerbar'** had been up for **9:00:59**.

`Daily' Graph (5 Minute Average)



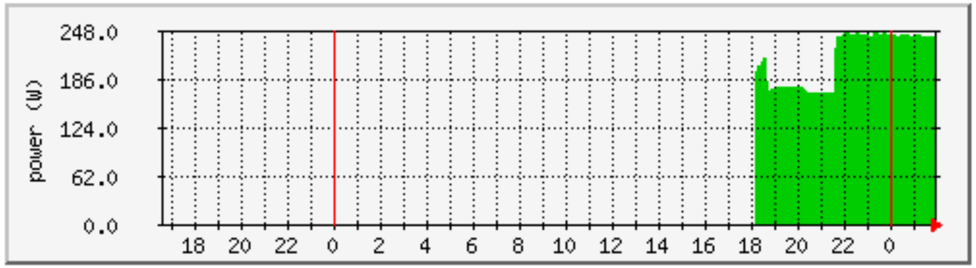
	Max	Average	Current
Power	256.6 Watt	224.3 Watt	253.9 Watt

Power Unit 1
Idle: 181 Watt
Streaming: 251 Watt

Total
Idle: 349 Watt
Streaming: 491 Watt
Difference: 142 Watt

The statistics were last updated **Friday, 1 October 2010 at 1:59**, at which time **'asd-powerbar'** had been up for **9:01:58**.

`Daily' Graph (5 Minute Average)



	Max	Average	Current
Power	244.3 Watt	211.2 Watt	239.2 Watt

Power Unit 2
Idle: 168 Watt
Streaming: 240 Watt

Conclusions

- 1 mainstream server is capable of sending 32.5 Gbit/s from disk to network
- SSD disks achieve high read performance, but filesystem is important (ZFS scales best)
- Saturating multiple 10GE NICs in 1 server is easy
- Large buffers are important
  - 9K MTU
  - Kernel max send and receive buffer set to 100MB
  - Application socket buffer set to 4.5 MB

Acknowledgements

- SARA: *Pieter de Boer, Freek Dijkstra, Igor Idziejczak, Tijs de Kler, Paul Melis, Hanno Pet, Peter Tavenier, Paul Wielinga*
- SURFnet: *Gerben van Malenstein, Erik-Jan Bos*
- Extreme Networks: *Rene Huntelerslag, Gihad Ghaibeh, Ramon Semmekrot, Trung Tran*
- Ciena: *Martin Bluethner, Jan Willem Elion, Kevin Mckernan, Harry Peng, David Yeung*
- CERN: *Edoardo Martelli*
- Leiden Observatory: *Simon Portegies Zwart*

- Partly funded by GigaPort3 and CineGrid.nl (EFRO, Pieken in de Delta, Provincie Noord-Holland, Gemeente Amsterdam)

Additional Information

- http://nrg.sara.nl/
- http://nrg.sara.nl/publications/RoN2010-D1.1.pdf
- http://nrg.sara.nl/publications/40G-Applications.pdf
- Email: nrg@sara.nl

Thank you

Ronald van der Pol

rvdp@sara.nl