Achieving the Science DMZ

Eli Dart, Brian Tierney and Eric Pouyoul, ESnet

Raj Kettimuthu, ANL

Jason Zurawski, Internet2

Guy Almes: Texas A&M

TIP2013, Honolulu Hawaii, January, 2013

Section 2: Bulk Data Transfer Tools and Data Transfer Nodes

Section Outline

Designing and Building a Data Transfer Nodes

- DTN Hardware Selection
- DTN Tuning

Data Transfer Tools

- Setting expectations
- What makes a fast data transfer tool
- Just say no to scp
- Open Source Tools
- Commercial Tools
- Tool Tuning

Designing and Building a Data Transfer Node

Eric Pouyoul and Brian Tierney, ESnet

Data Transfer Node

A DTN server is made of several subsystems. Each needs to perform optimally for the DTN workflow:

Storage: capacity, performance, reliability, physical footprint

Networking: protocol support, optimization, reliability

Motherboard: I/O paths, PCIe subsystem, IPMI

Chassis: adequate power supply, extra cooling

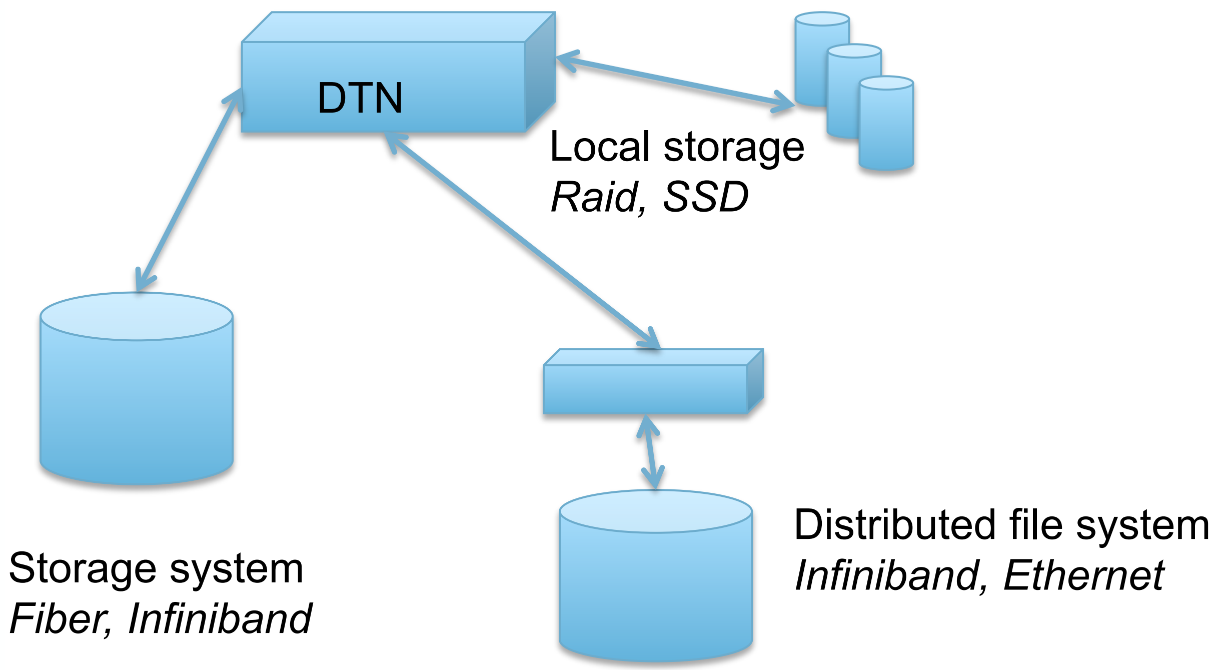
Note: the workflow we are optimizing for here is sequential reads/ write of large files, and a moderate number of high bandwidth flows.

We assume this host is dedicated to data transfer, and not doing data analysis/manipulation

Storage Architectures

There are multiple options for DTN Storage

- this does not really impact DTN node design



DTN Hardware Selection

The typical engineering trade-offs between cost, redundancy, performance, and so forth apply when deciding on what hardware to use for a DTN node

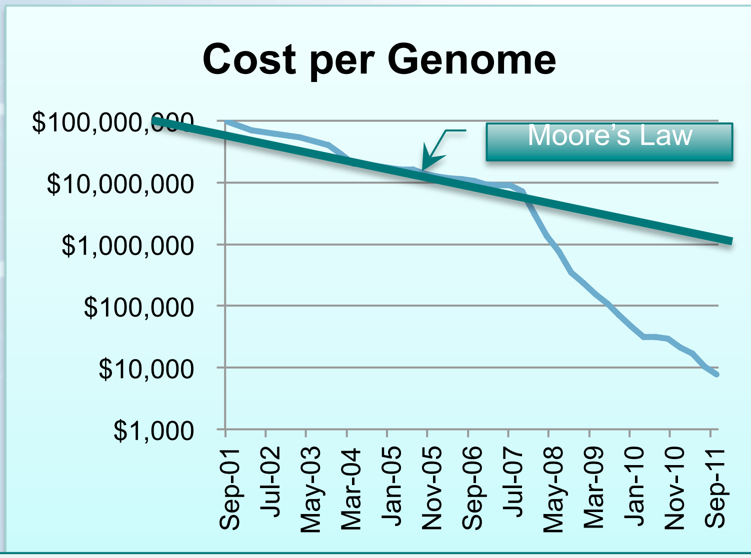
- e.g. redundant power supplies, SATA or SAS disks, cost/performance for RAID controllers, quality of NICs, etc.

We recommend getting a system that can be expanded to meet future storage/bandwidth requirements.

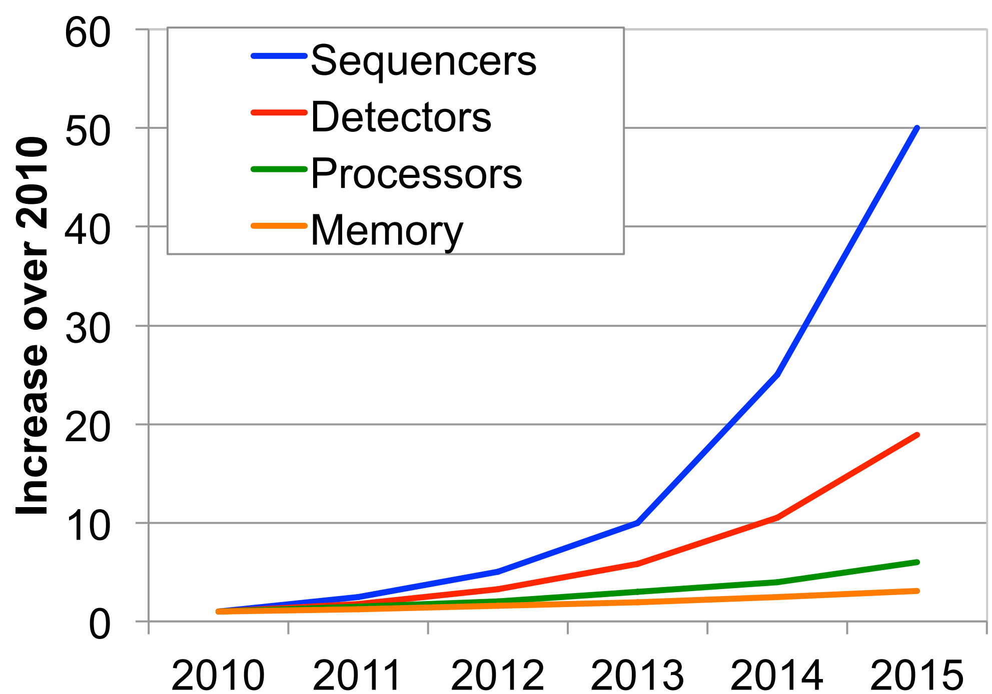
- science data is growing faster than Moore Law, so might as well get a host that can support 40GE now, even if you only have a 10G network

IMPORTANT: Buying the right hardware is not enough. Extensive tuning is needed for a well optimized DTN.

Sample Data Growth



Source: National Human Genome Research Institute



Motherboard and Chassis selection

Full 40GE requires a PCI Express gen 3 (aka PCIe gen3) motherboard

- this means are limited to a Intel Sandy Bridge or Ivy Bridge host.

Other considerations are memory speed, number of PCI slots, extra cooling, and an adequate power supply.

Intel's Sandy/Ivy Bridge (Ivy Bridge is about 20% faster) CPU architecture provides these features, which benefit a DTN:

- PCIe Gen3 support (up to 16 GB/sec)
- Turbo boost (up to 3.9 Ghz for the i7)
- Faster QPIC for communication between processors

Interrupt Binding Considerations

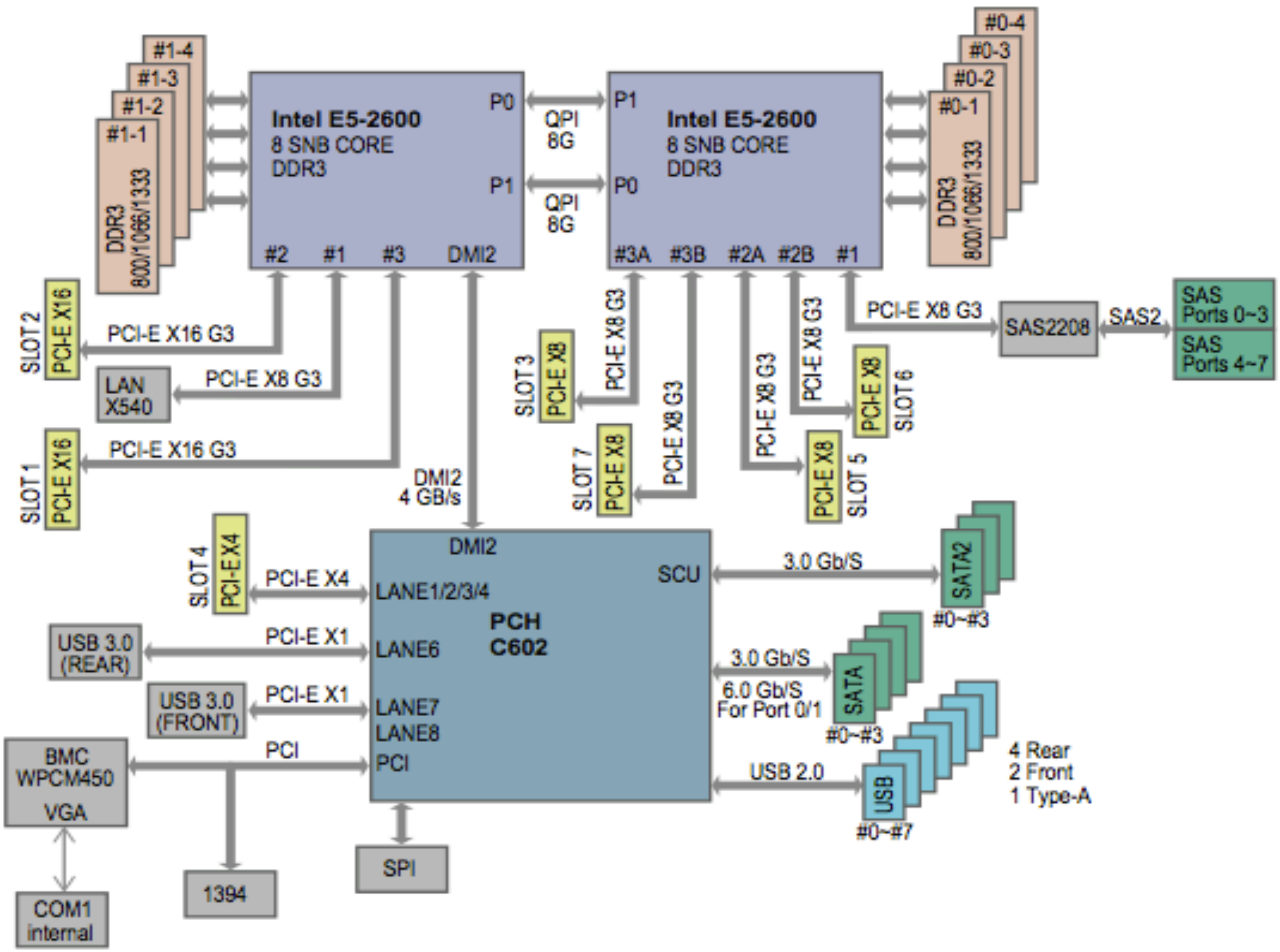
On a multi-processor host, your process might run on one processor, but your I/O interrupts on another processor.

- When this happens there will be a huge performance hit.
  - a single 10G NIC can saturate the QPIC that connects the two processors.
- You may want to consider getting a single CPU system, and avoid dual-processor motherboards if you only need 3 PCIe slots.
- If you need to optimize for a small number of very fast (> 6Gbps) flows, this means you will need to manage IQR bindings by hand.
  - If you are optimizing for many 500Mbps-1Gbps flows, this will be less of an issue.
  - You may be better off doing 4x10GE instead of 1x40GE, as you will have more control mapping IRQs to processors.

The impact of this is even greater with Sandy/Ivy Bridge Hosts, as the PCI bus slots are connected directly to a processor.

Intel Sandy/Ivy Bridge

SUPERO X9DAX-iF/-7F/-iTF/-7TF Motherboard User's Manual



Memory and PCI Slot Considerations

- Memory
  - We recommend 32GB of RAM for a DTN node. More is better.
- PCI Slots
  - Be sure to get the right number of the right type of PCI slots for your needs.
  - PCI slots are defined by:
    - Form factor: This is the length of the slots, referred as the number of PCI lanes it can support. A 16 lane controller is twice as long as a 8 lane controller.
    - Number of wired lanes: not all lanes of the slot may be wired. Some 8 lanes controller may only have 4 lanes wired.
    - PCIe 2.0 is 500 MB/sec per lane. A typical host supports 8 lane (x8) cards, or up to 4 GB/sec. A high-end host might have 16 lane (x16) slots, or up to 8 GB/sec.
    - PCIe 3.0 doubles this bandwidth.

PCI Bus Considerations

Make sure the motherboard you select has the right number of slots with the right number of lanes for you planned usage. For example:

- 10GE NICs require a 8 lane PCIe-2 slot
- 40G/QDR NICs require a 8 lane PCIe-3 slot
- HotLava has a 6x10GE NIC that requires a 16 lane PCIe-2 slot
- Most RAID controllers require 8 lane PCIe-2 slot
- Very high-end RAID controllers for SSD might require a 16 lane PCIe-2 slot or a 8 line PCIe-3 slot
- A high-end Fusion IO ioDrive requires a 16 lane PCIe-2 slot

Some motherboards to look at include the following:

- SuperMicro X9DR3-F
- Sample Dell Server (Poweredge r320-r720)
- Sample HP Server (ProLiant DL380p gen8 High Performance model)

Storage Subsystem Selection

- Deciding what storage to use in your DTN is based on deciding what you are optimizing for:
  - performance, reliability, and capacity, and cost.
- SATA disks historically have been cheaper and higher capacity, while SAS disks typically have been the fastest.
- However these technologies have been converging, and with SATA 3.1 is less true.

SSD Storage: on its way to becoming mainstream.

SSD

SSD storage costs much more than traditional hard drives (HD), but are much faster. They come in different styles:

- PCIe card: some vendors (Fusion I/O) build PCI cards with SSD.
  - These are the fastest type of SSD: up to several GBytes/sec per card.
    - Note that this type of SSD is typically not hot-swapable.
- HD replacement: several vendors now sell SSD-based drives that have the same form factor as traditional drives such as SAS and SATA.
  - The downside to this approach is that performance is limited by the RAID controller, and not all controllers work well with SSD.
    - Be sure that your RAID controller is “SSD capable”.
- Note that the price of SSD is coming down quickly, so a SSD-based solution may be worth considering for your DTNs.

RAID Controllers

- Often optimized for a given workload, rarely for performance.
- RAID0 is the fastest of all RAID levels but is also the least reliable.
- The performance of the RAID controller is a factor of the number of drives and its own processing engine.

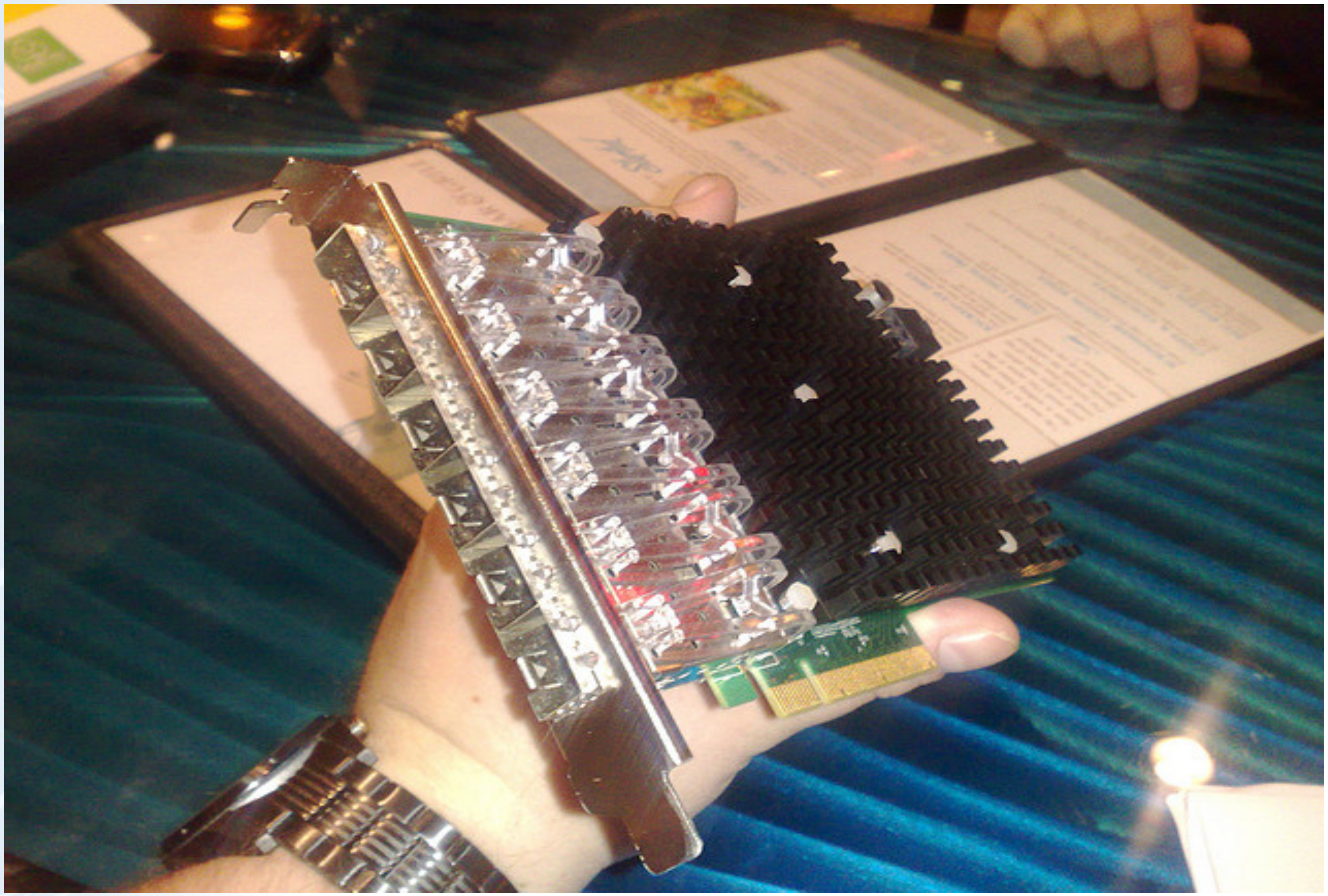
RAID Controller

Be sure your RAID controller has the following:

- 1GB of on-board cache
- PCIe Gen3 support
- dual-core RAID-on-Chip (ROC) processor if you will have more than 8 drives

One example of a RAID card that satisfies these criteria is the Areca ARC-1882i.

Networking Subsystem



Network Subsystem Selection

There is a huge performance difference between cheap and expensive 10G NICs.

- You should not go cheap with the NIC, as a high quality NIC is important for an optimized DTN host.

NIC features to look for include:

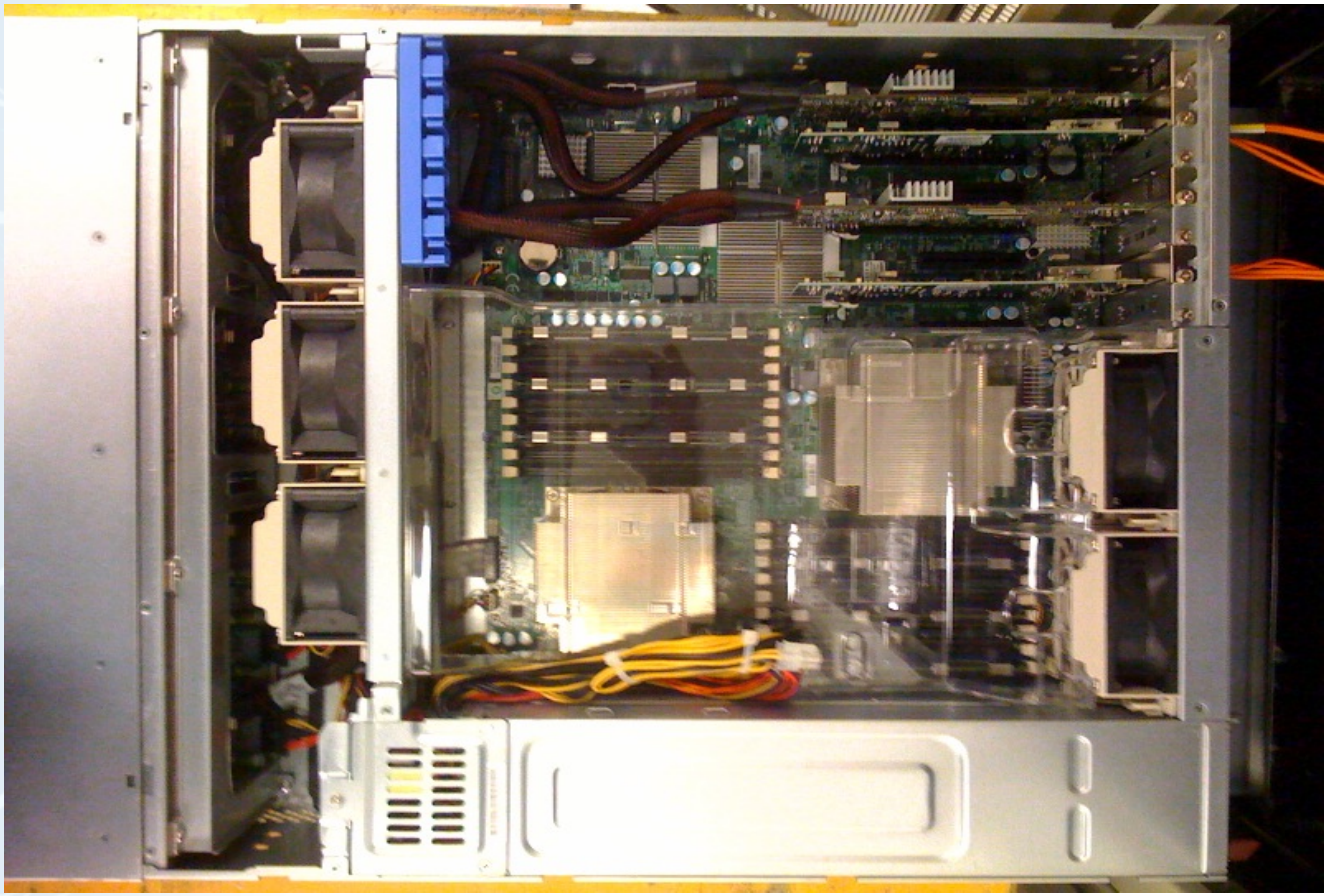
- support for interrupt coalescing
- support for MSI-X
- TCP Offload Engine (TOE)
- support for zero-copy protocols such as RDMA (RoCE or iWARP)

Note that many 10G and 40G NICs come in dual ports, but that does not mean if you use both ports at the same time you get double the performance. Often the second port is meant to be used as a backup port.

- True 2x10G capable cards include the Myricom 10G-PCIE2-8C2-2S and the Mellanox MCX312A-XCBT.

Several vendors now support iWarp, but currently Mellanox is the only vender that supports RoCE in hardware.

Tuning the Data Transfer Host



Tuning

Defaults are usually not appropriate for performance.

What needs to be tuned:

- BIOS
- Firmware
- Device Drivers
- Networking
- File System
- Application

DTN Tuning

Tuning your DTN host is extremely important. We have seen overall IO throughput of a DTN more than double with proper tuning.

Tuning can be as much art as a science. Due to differences in hardware, its hard to give concrete advice.

Here are some tuning settings that we have found do make a difference.

- This tutorial assumes you are running a Redhat-based Linux system, but other types of Unix should have similar tuning nobs.

Note that you should always use the most recent version of the OS, as performance optimizations for new hardware are added to every release.

Network Tuning

```
# add to /etc/sysctl.conf
net.core.rmem_max = 33554432
net.core.wmem_max = 33554432
net.ipv4.tcp_rmem = 4096 87380 33554432
net.ipv4.tcp_wmem = 4096 65536 33554432
net.core.netdev_max_backlog = 250000

Add to /etc/rc.local
# increase txqueuelen
/sbin/ifconfig eth2 txqueuelen 10000
/sbin/ifconfig eth3 txqueuelen 10000
```

```
# make sure cubic and htcp are loaded
/sbin/modprobe tcp_htcp
/sbin/modprobe tcp_cubic
# set default to CC alg to htcp
net.ipv4.tcp_congestion_control=htcp

# with the Myricom 10G NIC increasing
    interrupt coalencing helps a lot:
/usr/sbin/ethtool -C ethN rx-usecs 75
```

BIOS Tuning

For PCI gen3-based hosts, you should

- enable "turbo boost",
- disable hyperthreading and node interleaving.

More information on BIOS tuning is described in this document:

http://www.dellhpcsolutions.com/assets/pdfs/Optimal_BIOS_HPC_Dell_12G.v1.0.pdf

I/O Scheduler

The default Linux scheduler is the "fair" scheduler. For a DTN node, we recommend using the "deadline" scheduler instead.

To enable deadline scheduling, add "elevator=deadline" to the end of the "kernel' line in your /boot/grub/grub.conf file, similar to this:

```
kernel /vmlinuz-2.6.35.7 ro root=/dev/VolGroup00/
LogVol00 rhgb quiet elevator=deadline
```

File System Tuning

We recommend using the ext4 file system in Linux for DTN nodes.

Increasing the amount of "readahead" usually helps on DTN nodes where the workflow is mostly sequential reads.

- However you should test this, as some RAID controllers do this already, and changing this may have adverse affects.

Setting readahead should be done at system boot time. For example, add something like this to /etc/rc.local:

```
/sbin/blockdev --setra 262144 /dev/sdb
```

More information on readahead:

http://www.kernel.org/doc/ols/2004/ols2004v2-pages-105-116.pdf

EXT4 Tuning

The file system should be tuned to the physical layout of the drives.

- Stride and stripe-width are used to align the volume according to the stripe-size of the RAID.
  - stride is calculated as Stripe Size / Block Size.
  - stripe-width is calculated as Stride * Number of Disks Providing Capacity.

Disabling journaling will also improve performance, but reduces reliability.

Sample mkfs command:

```
/sbin/mkfs.ext4  /dev/sdb1 –b 4096 –E stride=64 \
stripe-width=768 –O ^has_journal
```

File System Tuning (cont.)

There are also tuning settings that are done at mount time. Here are the ones that we have found improve DTN performance:

- data=writeback
  - this option forces ext4 to use journaling only for metadata. This gives a huge improvement in write performance
- inode_readahead_blks=64
  - this specifies the number of inode blocks to be read ahead by ext4's readahead algorithm. Default is 32.
- Commit=300
  - this parameter tells ext4 to sync its metadata and data every 300s. This reduces the reliability of data writes, but increases performance.
- noatime,nodiratime
  - these parameters tells ext4 not to write the file and directory access timestamps.

Sample fstab entry:

```
/dev/sdb1 /storage/data1 ext4
inode_readahead_blks=64,data=writeback,barrier=0,commit=300,noa
time,nodiratime
```

For more info see: http://kernel.org/doc/Documentation/filesystems/ext4.txt

RAID Controller

Different RAID controllers provide different tuning controls. Check the documentation for your controller and use the settings recommended to optimize for large file reading.

- You will usually want to disable any "smart" controller built-in options, as they are typically designed for different workflows.

Here are some settings that we found increase performance on a 3ware RAID controller. These settings are in the BIOS, and can be entered by pressing Alt+3 when the system boots up.

- Write cache – Enabled
- Read cache – Enabled
- Continue on Error – Disabled
- Drive Queuing – Enabled
- StorSave Profile – Performance
- Auto-verify – Disabled
- Rapid RAID recovery – Disabled

Virtual memory Subsystem

Setting dirty_background_bytes and dirty_bytes improves write performance.

For our system, the settings that gave best performance was:

```
echo 1000000000 > /proc/sys/vm/dirty_bytes
echo 1000000000 > /proc/sys/vm/dirty_background_bytes
```

For more information see:

http://www.kernel.org/doc/Documentation/sysctl/vm.txt

SSD Issues

Tuning your SSD is more about reliability and longevity than performance, as each flash memory cell has a finite lifespan that is determined by the number of "program and erase (P/E)" cycles. Without proper tuning, SSD can die within months.

- **never** do "write" benchmarks on SSD: this will damage your SSD quickly.

Modern SSD drives and modern OSes should all includes TRIM support, which is important to prolong the live of your SSD. Only the newest RAID controllers include TRIM support (late 2012).

Swap

To prolong SSD lifespan, do not swap on an SSD. In Linux you can control this using the sysctl variable vm.swappiness. E.g.: add this to /etc/sysctl.conf:

  – vm.swappiness=1

This tells the kernel to avoid unmapping mapped pages whenever possible.

Avoid frequent re-writing of files (for example during compiling code from source), use a ramdisk file system (tmpfs) for /tmp /usr/tmp, etc.

ext4 file system tuning for SSD

These mount flags should be used for SSD partitions.

- noatime: Reading accesses to the file system will no longer result in an update to the atime information associated with the file. This eliminates the need for the system to make writes to the file system for files which are simply being read.
- discard: This enables the benefits of the TRIM command as long for kernel version >=2.6.33.

Sample /etc/fstab entry:

```
/dev/sda1 /home/data ext4 defaults,noatime,discard 0 1
```

For more information see:

http://wiki.archlinux.org/index.php/Solid_State_Drives

Benchmarking

Simulating a single thread writing sequentially a file can be done using dd as:

```
$ dd if=/dev/zero of=/storage/data1/file1 bs=4k
  count=33554432
```

Simulating sequentially reading a file is done by:

```
$ dd if=/storage/data1/file1 of=/dev/null bs=4k
```

Sample Hardware configuration

Motherboard: SuperMicro X9DR3-F

CPU: 2 x Intel(R) Xeon(R) CPU E5-2667 0 @ 2.90GHz

Memory: 64G (8 x 8GiB Kingston DDR3 1600 MHz / 0.6 ns)

RAID: 2 x 3Ware 9750SA-24i (24 ports SAS) connected to CPU 0

- Areca ARC-1882i should be even better, but we have not tested it

Network Controller: Myricom 10G-PCIE2-8C2-2S

- Use Mellanox cards if you are interested in RDMA

Disks: 16 x Seagate 1TB SAS HDD 7,200 RPM drives

Total cost: $12K in mid 2012

Bulk Data Transfer Tools

Time to Copy 1 Terabyte

10 Mbps network : 300 hrs (12.5 days)

100 Mbps network : 30 hrs

1 Gbps network : 3 hrs (are your disks fast enough?)

10 Gbps network : 20 minutes (need really fast disks and filesystem)

These figures assume some headroom left for other users

Compare these speeds to:

- USB 2.0 portable disk
  - 60 MB/sec (480 Mbps) peak
  - 5-15 MB/sec reported typical performance
  - 15-40 hours to load 1 Terabyte

Sample Data Transfer Results

Using the right tool is very important

Sample Results: Berkeley, CA to Argonne, IL (near Chicago). RTT = 53 ms, network capacity = 10Gbps.

Tool	Throughput
– scp:	140 Mbps
– HPN patched scp:	1.2 Gbps
– ftp	1.4 Gbps
– GridFTP, 4 streams	5.4 Gbps
– GridFTP, 8 streams	6.6 Gbps

- Note that to get more than 1 Gbps (125 MB/s) disk to disk requires RAID.

Data Transfer Tools

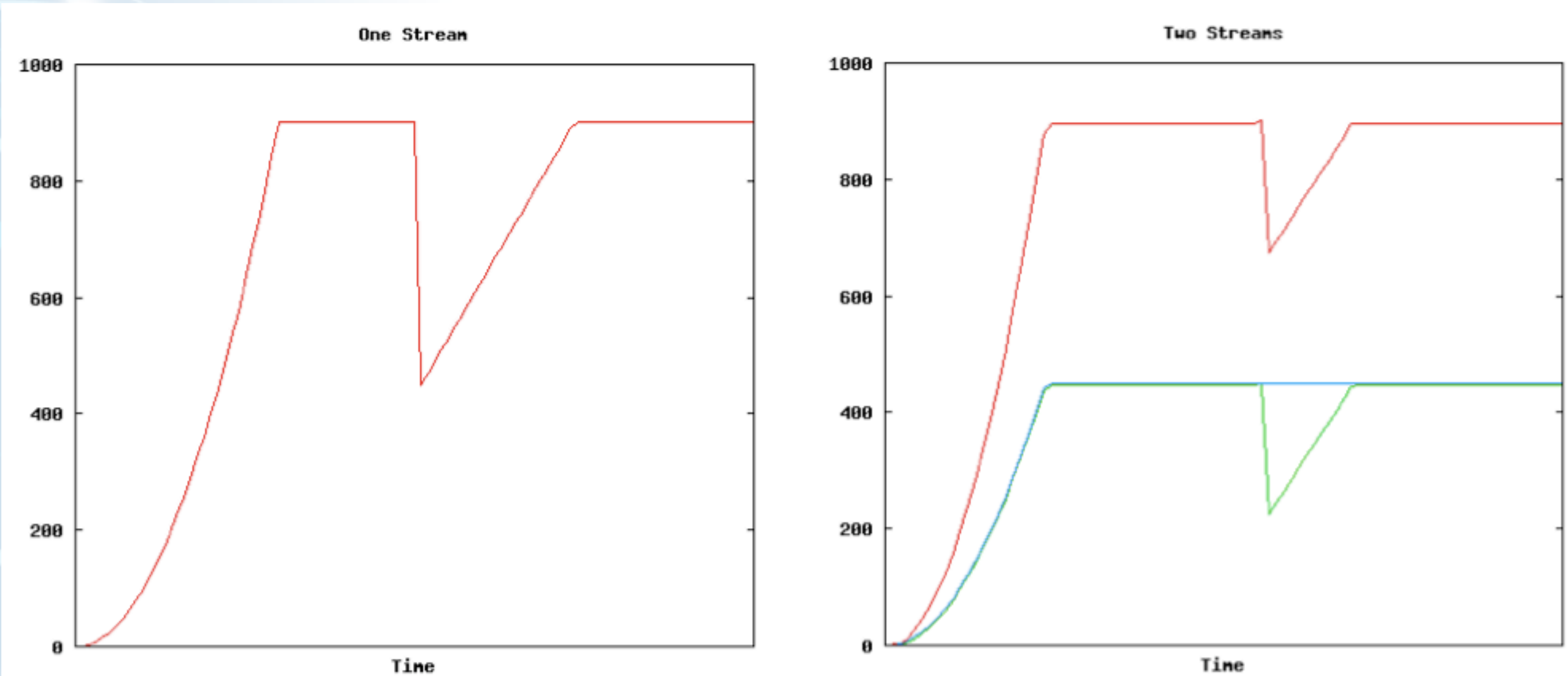
Parallelism is key

- It is much easier to achieve a given performance level with four parallel connections than one connection
- Several tools offer parallel transfers

Latency interaction is critical

- Wide area data transfers have much higher latency than LAN transfers
- Many tools and protocols assume a LAN
- Examples: SCP/SFTP, HPSS mover protocol

Parallel Streams Help With TCP Congestion Control Recovery Time



Why Not Use SCP or SFTP?

Pros:

- Most scientific systems are accessed via OpenSSH
- SCP/SFTP are therefore installed by default
- Modern CPUs encrypt and decrypt well enough for small to medium scale transfers
- Credentials for system access and credentials for data transfer are the same

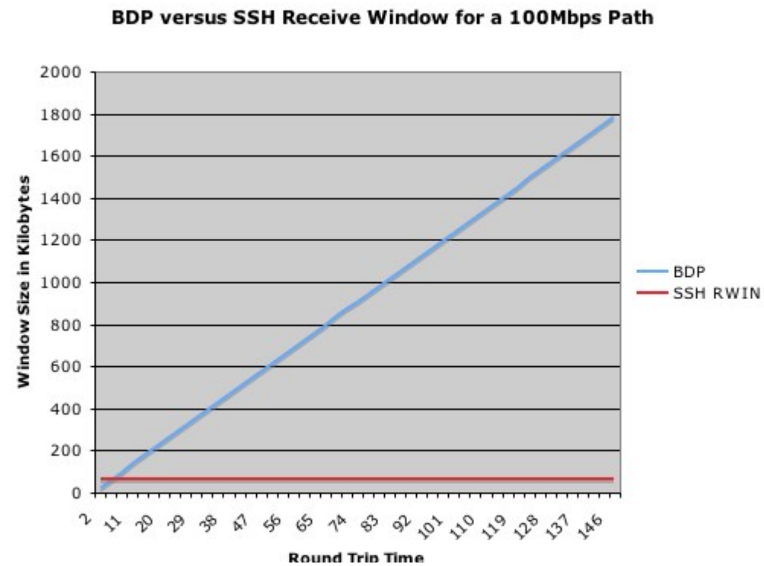
Cons:

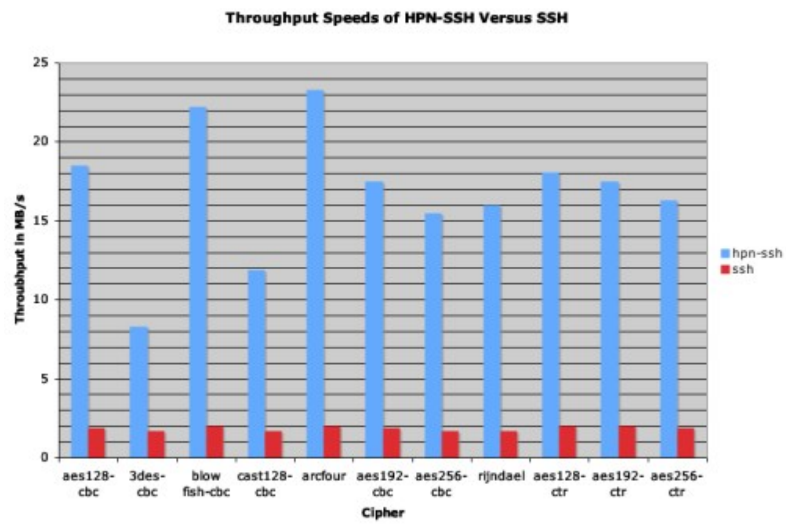
- The protocol used by SCP/SFTP has a fundamental flaw that limits WAN performance
- CPU speed doesn't matter – latency matters
- Fixed-size buffers reduce performance as latency increases
- It doesn't matter how easy it is to use SCP and SFTP – they simply do not perform

Verdict: Do Not Use Without Performance Patches

A Fix For scp/sftp

- PSC has a patch set that fixes problems with SSH
    - http://www.psc.edu/networking/projects/hpn-ssh/
- Significant performance increase
- Advantage – this helps rsync too





sftp

Uses same code as scp, so don't use sftp WAN transfers unless you have installed the HPN patch from PSC

But even with the patch, SFTP has yet another flow control mechanism

- By default, sftp limits the total number of outstanding messages to 16 32KB messages.
- Since each datagram is a distinct message you end up with a 512KB outstanding data limit.
- You can increase both the number of outstanding messages ('-R') and the size of the message ('-B') from the command line though.

Sample command for a 128MB window:

- sftp -R 512 -B 262144 user@host:/path/to/file outfile

FDT

FDT = Fast Data Transfer tool from Caltech

- http://monalisa.cern.ch/FDT/
- Java-based, easy to install
- used by US-CMS project
- being deployed by the DYNES project

GridFTP

GridFTP from ANL has features needed to fill the network pipe

- Buffer Tuning
- Parallel Streams

Supports multiple authentication options

- Anonymous
- ssh
- X509

Ability to define a range of data ports

- helpful to get through firewalls

New Partnership with ESnet and Globus Online to support Globus Online for use in Science DMZs

Much more detail from Raj in the next section of the tutorial.

Other Data Transfer Tools

bbcp: http://www.slac.stanford.edu/~abh/bbcp/

- supports parallel transfers and socket tuning
- bbcp -P 4 -v -w 2M myfile remotehost:filename

lftp: http://lftp.yar.ru/

- parallel file transfer, socket tuning, HTTP transfers, and more.
- lftp -e 'set net:socket-buffer 4000000; pget -n 4 [http|ftp]://site/path/file; quit'

axel: http://axel.alioth.debian.org/

- simple parallel accelerator for HTTP and FTP.
- axel -n 4 [http|ftp]://site/file

Commercial Data Transfer Tools

There are several commercial UDP-based tools

- Aspera: http://www.asperasoft.com/
- Data Expedition: http://www.dataexpedition.com/
- TIXstream: http://www.tixeltec.com/tixstream_en.html

These should all do better than TCP on a congested, high-latency path

- advantage of these tools less clear on an uncongested path

They all have different, fairly complicated pricing models

Next Generation Tools/Protocols

RDMA-based tools:

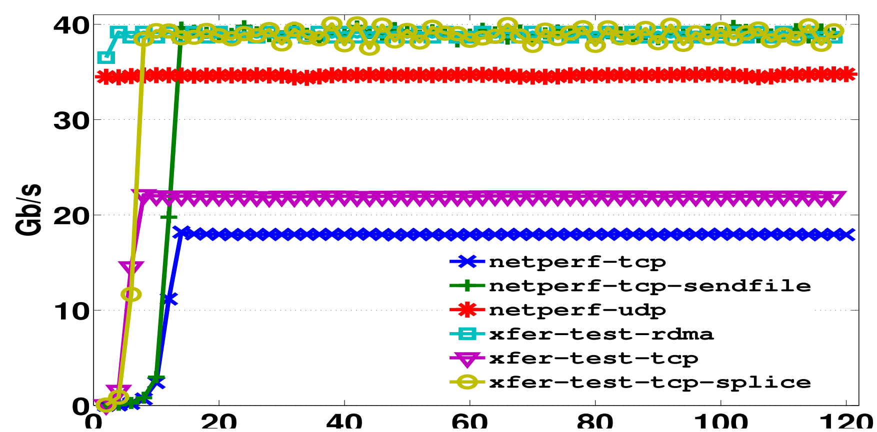
- Several groups have been experimenting with RDMA over the WAN
  - XIO driver for GridFTP (IU, OSU)
  - RFTP: BNL
- Over a dedicated layer-2 circuit, performance is the same as TCP, with **much** less CPU
- Requires hardware support on the NIC (e.g.: Mellanox)
  - Software version exists, but requires custom kernel and is slower
- RDMA tuning can be quite tricky to get right

Session Layer Networking / Phoebus:

Phoebus Gateway can by used to translate being the LAN protocol (e.g. TCP) and a more efficient WAN protocol (e.g.: RDMA)

Recent Testbed Results: Single flow 40G

Tool	Protocol	Gbps	Send CPU	Recv CPU
netperf	TCP TCP-sendfile UDP	17.9 39.5 34.7	100% 34% 100%	87% 94% 95%
xfer_test	TCP TCP-splice RoCE	22 39.5 39.2	100% 43% 2%	91% 91% 1%
GridFTP	TCP UDT RoCE	13.3 3.6 13	100% 100% 100%	94% 100% 150%



Tuning your Data Transfer Tools

Be sure to check the following:

- What is your host's maximum TCP window size?
  - 32M is good for most many environments
  - More for jumbo frames or very long RTT paths
- Which TCP congestion algorithm are you using?
  - Cubic or HTCP are usually best

- How many parallel streams are you using?
  - Use as few as possible that fill the pipe, usually 2-4 streams
  - Too many streams usually end up stepping on each other
    - May need more streams in cases of:
      - Very high RTT paths
      - Traversing slow firewalls
      - Paths without enough switch buffering

Questions?

Thanks!

Eric Pouyoul: lomax@es.net

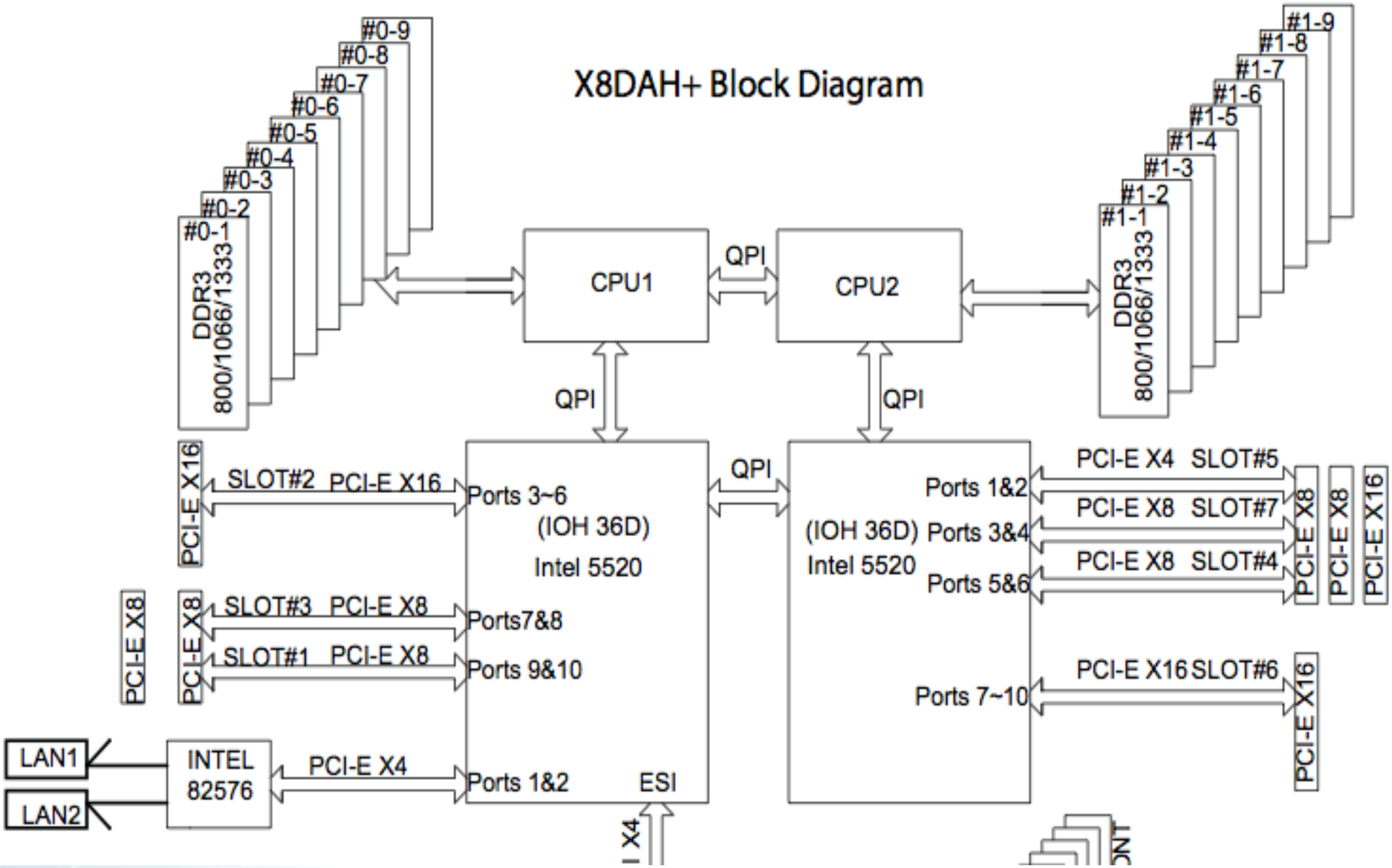
Brian Tierney, bltierney@es.net

http://www.es.net/

http://fasterdata.es.net/

Extra Slides

Intel PCIe Gen2: SuperMicro X8DAH+-F



Interrupt Affinity

- Interrupts are triggered by I/O cards (storage, network). High performance means lot of interrupts per seconds
- Interrupt handlers are executed on a core
- Depending on the scheduler, core 0 gets all the interrupts, or interrups are dispatched in a round-robin fashion among the cores: both are bad for performance:
    - Core 0 get all interrupts: with very fast I/O, the core is overwhelmed and becomes a bottleneck
    - Round-robin dispatch: very likely the core that executes the interrupt handler will not have the code in its L1 cache.
    - Two different I/O channels may end up on the same core.

A simple solution: interrupt binding

- Each interrupt is statically bound to a given core (network -> core 1, disk -> core 2)
- Works well, but can become an headache and does not fully solve the problem: one very fast card can still overwhelm the core.
- Needs to bind application to the same cores for best optimization: what about multi-threaded applications, for which we want one thread = one core ?

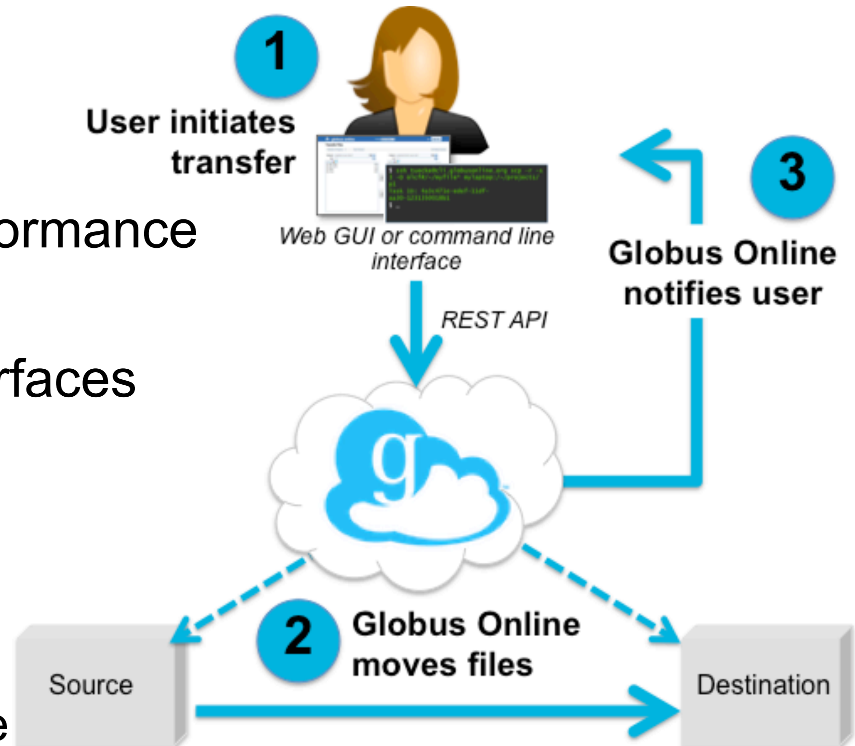
globus online

Reliable, high-performance file transfer… made easy.

Raj Kettimuthu
Argonne National Laboratory and University of Chicago

What is Globus Online?

- **Move, sync files**
  - Easy "fire-and-forget" transfers
  - Automatic fault recovery & high performance
  - Across multiple security domains
  - Web, command line, and REST interfaces
- **Minimize IT costs**
  - Software as a Service (SaaS)
    - No client software installation
    - New features automatically available
  - Consolidated support & troubleshooting
  - Simple endpoint installation with Globus Connect and GridFTP
- **Recommended by XSEDE, Blue Waters, NERSC, ALCF, NCAR, ESnet, many Universities**



LIVE DEMO

globus online

Enabling your data transfer node as a Globus Online Endpoint

Globus Connect Multi User

Download from http://www.globusonline.org/gcmu

tar xzf globusconnect-multiuser-latest.tgz

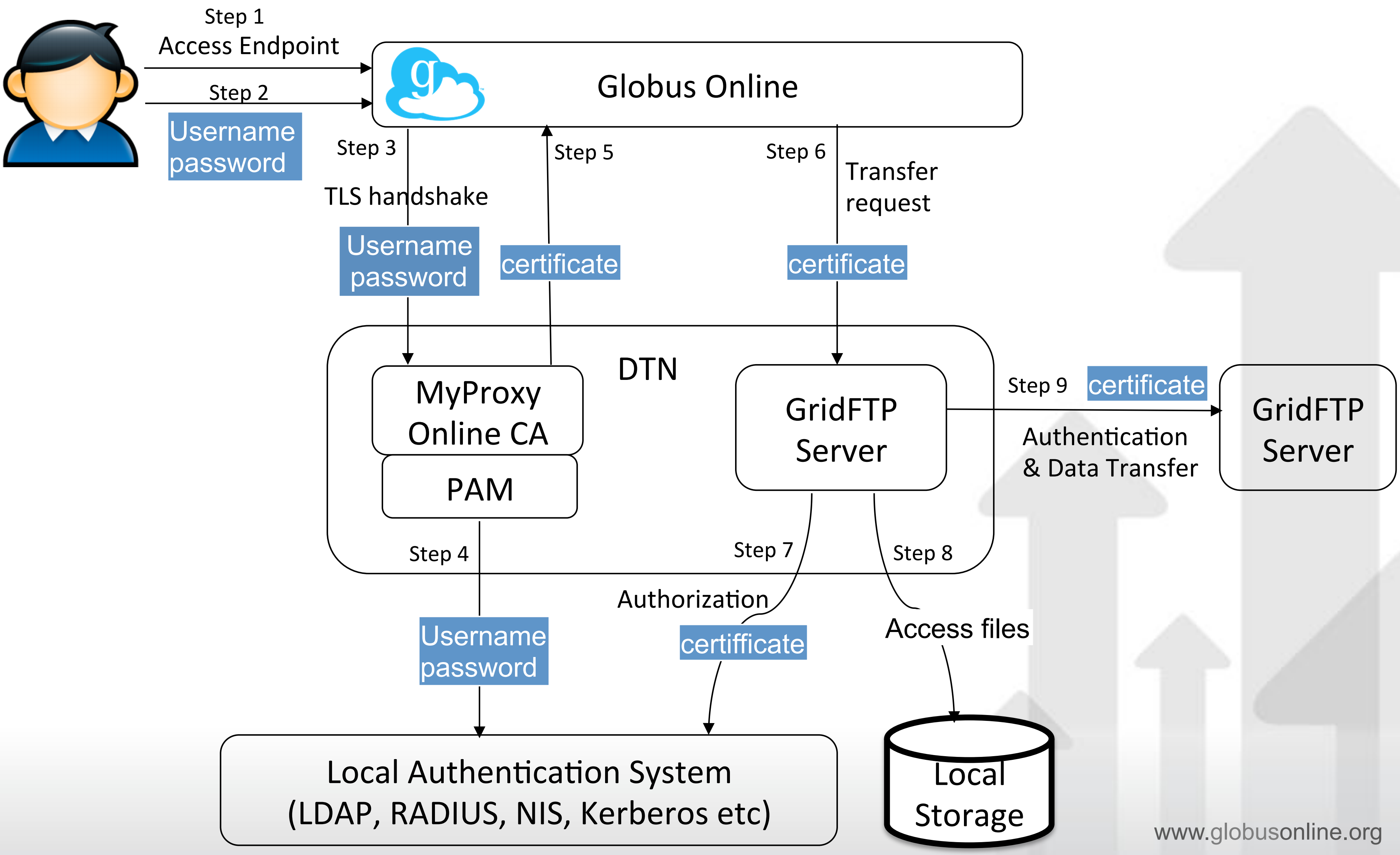
cd gcmu-1.1.3/

./install

Follow the prompts…

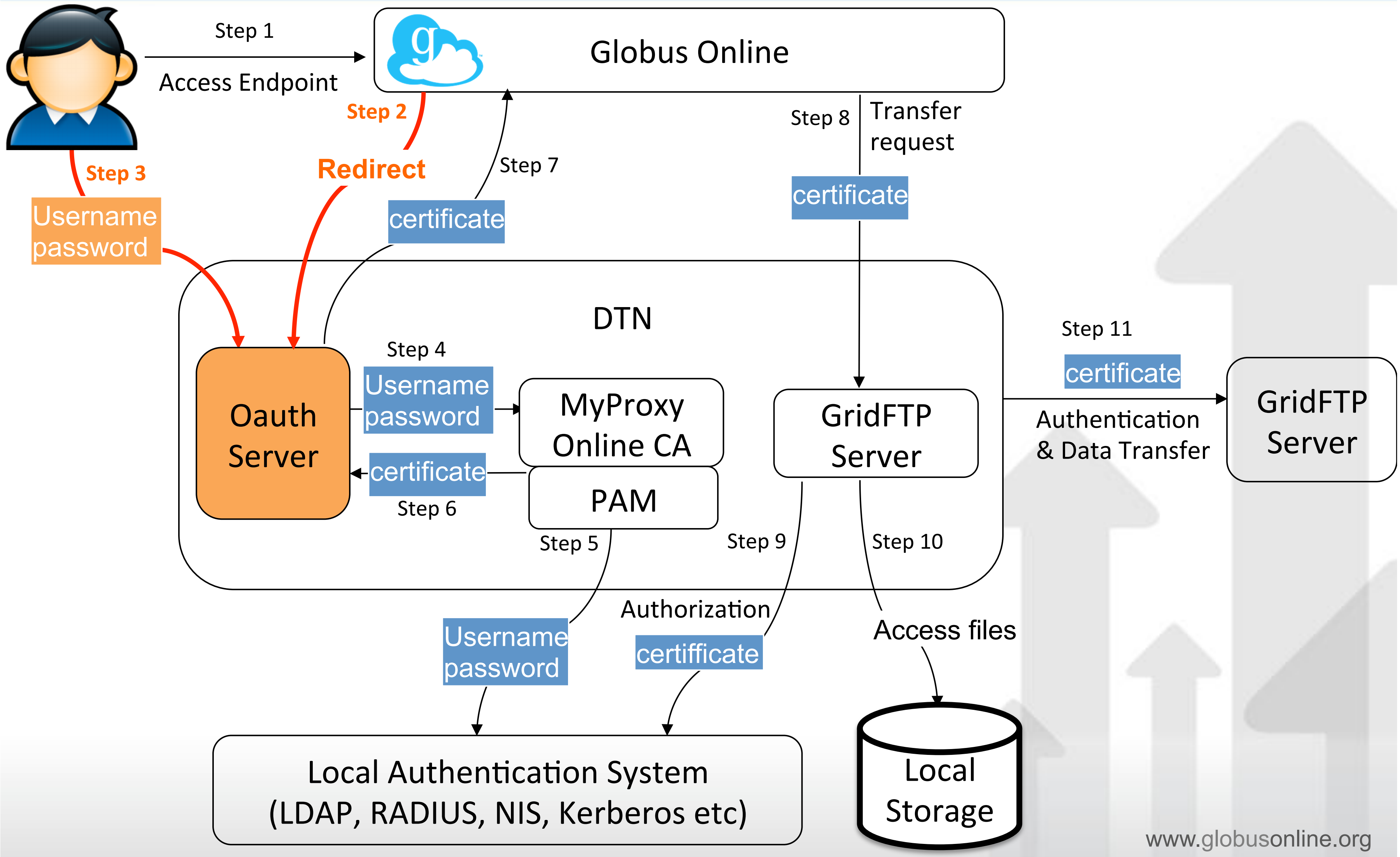
Globus Connect Multi-User Demo

Globus Connect Multi User



www.globusonline.org

… with MyProxy OAuth



www.globusonline.org

GCMU Deployments

Questions?

For More Information

- **Visit https://www.globusonline.org/signup to:**
  - Get a free account and start moving files
- **Visit www.globusonline.org for:**
  - Tutorials, FAQs, Pro Tips, Troubleshooting
  - Papers
  - Case Studies
- **Contact support@globusonline.org for:**
  - Help getting started
  - Help using the service
- **Follow us at @globusonline on Twitter and Globus Online on Facebook**

globus online

Extra Slides

GCMU makes GridFTP setup trivial

- **Make GridFTP deployment trivial**
- GridFTP transfers can be achieved “instantly” even by non-experts
- **Automate the process of configuring security**
- Avoid the need for any end-user or system administrator involvement in security configuration
- **Reduce burden on both users and administrators**
- Eliminate frequent sources of errors in GridFTP configuration and use.

Details

- **GCMU may require firewall configuration**
  - Inbound ports: 2811, 7512, and 50,000-51,000
  - Outbound ports: 50,000-51,000
- **GCMU GridFTP cannot be (easily) used with other GridFTP clients besides GO**
  - GCMU uses a GO-issued cert, not a host cert
- **GridFTP has many configuration options**
  - E.g., Gridmap file can be used to allow other certs
  - Limit access to particular directories
  - You must re-apply these manually after an update
- **MyProxy CA default proxy lifetime is 12 hours**