Distributed IaaS Clouds and 100G Networking for HEP applications

HPCS 2013 June 3

Ian Gable

A.Agarwal, A.Charbonneau, C.Leavett-Brown, K Lewall. R. Impey , M.Paterson,
W. Podiama, R.J. Sobie, R.Taylor
University of Victoria, NRC

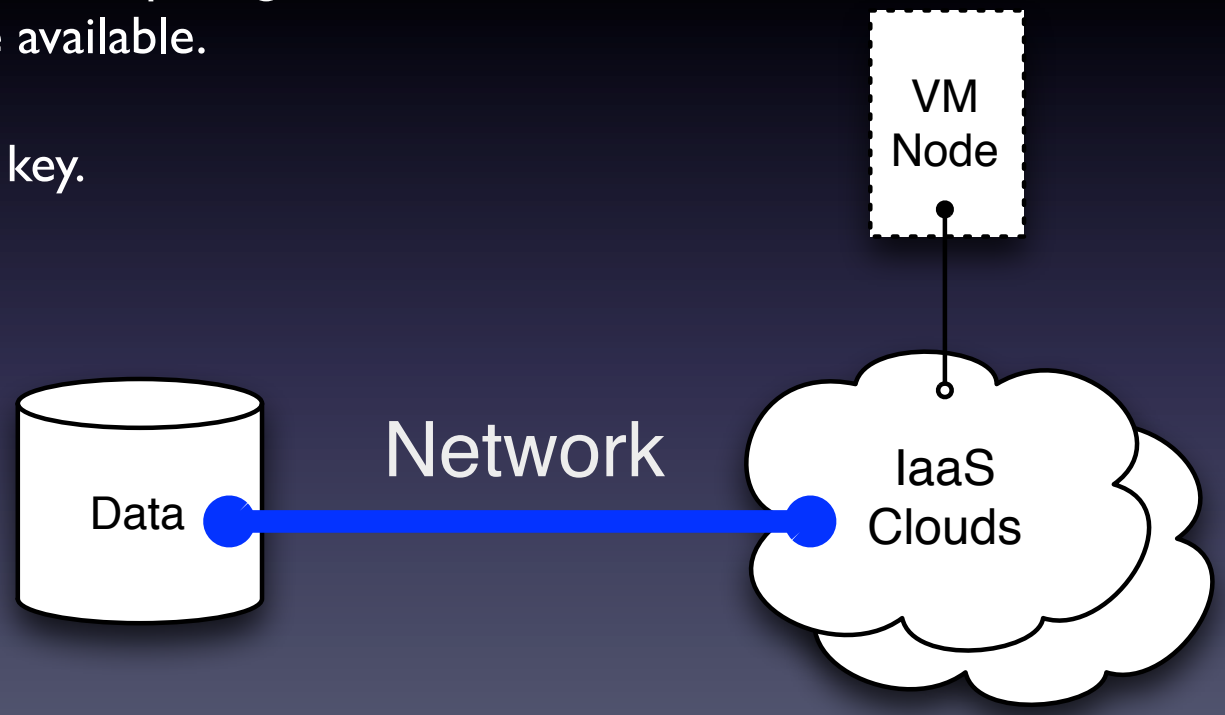
M. Hay, D. McWilliam Y. Savard, T. Tam
BCNet and CANARIE

A. Barczyk, R. Hockett, D. Kcira, I. Legrand, S. McKee, A. Mughal,
H. Newman, S. Rosza, S. Timoteo, R. Voicu
Caltech and University of Michigan

Typically, computing tasks have been moved to the data.

With Infrastructure-as-a-Service clouds we want to move the data to computing on demand as resources become available.

Networks are the key.



Outline

Running ATLAS jobs on distributed IaaS cloud

- Motivation
- Software required
- Results from 14 months of production operation

100G network demonstrations at Super Computing 2012

- HEP networks today
- Physical setup
- Results

ATLAS Experiment in One Slide

- A particle detector at CERN which gathers a vast amount of data (100+ PB in last 2 years)
- Analysis and simulation computational tasks (i.e. jobs) are embarrassingly parallel
- Most of the computation is completed today on the WLCG Grid

See Reda Tafirout's talk:
"ATLAS Experiment: Big Data + Big Compute = Big Discovery"
Immediately following this talk

Many facilities are evolving into clouds

Can we use these sites in a federated way?

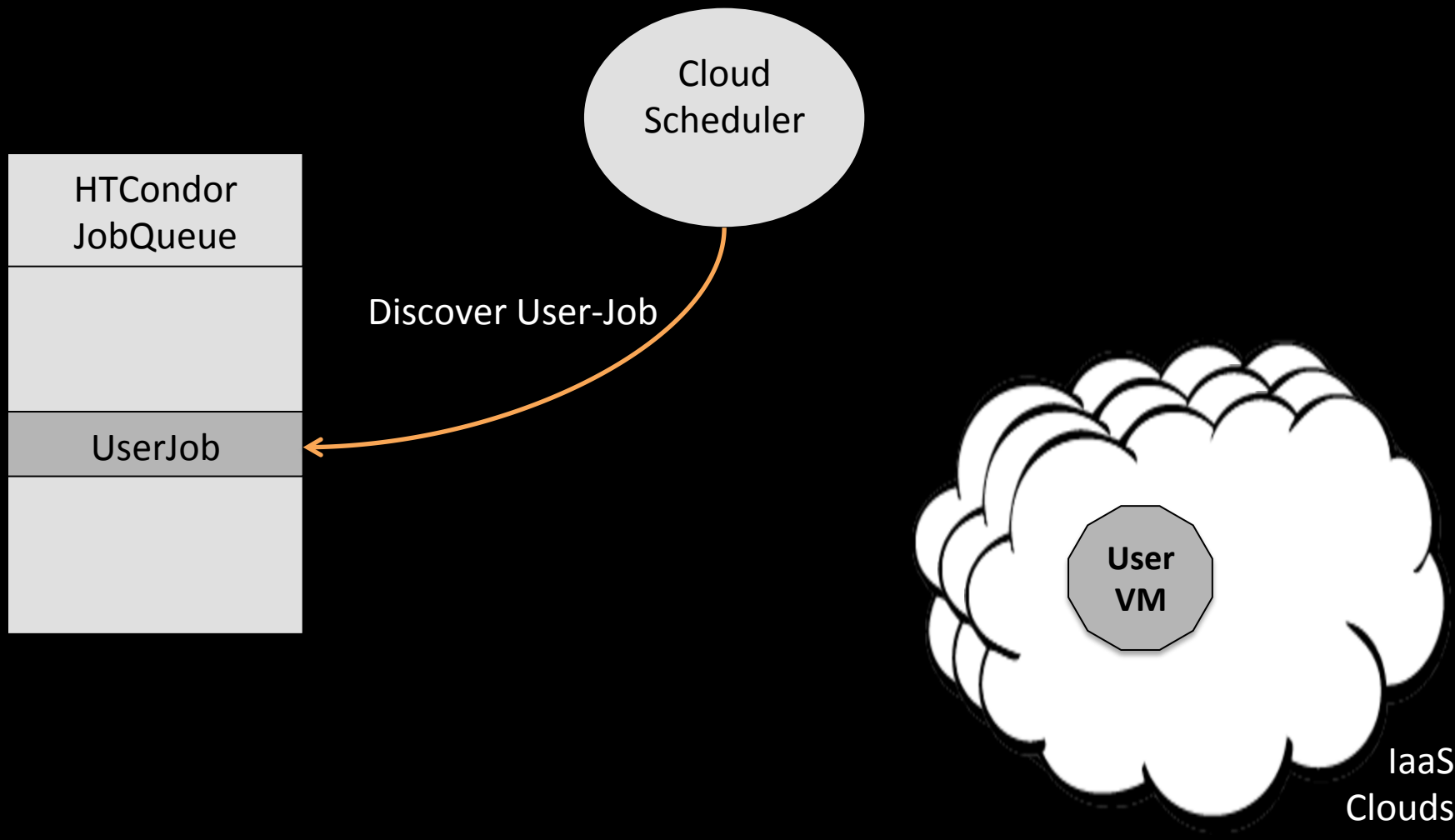
And integrate them into our existing systems?

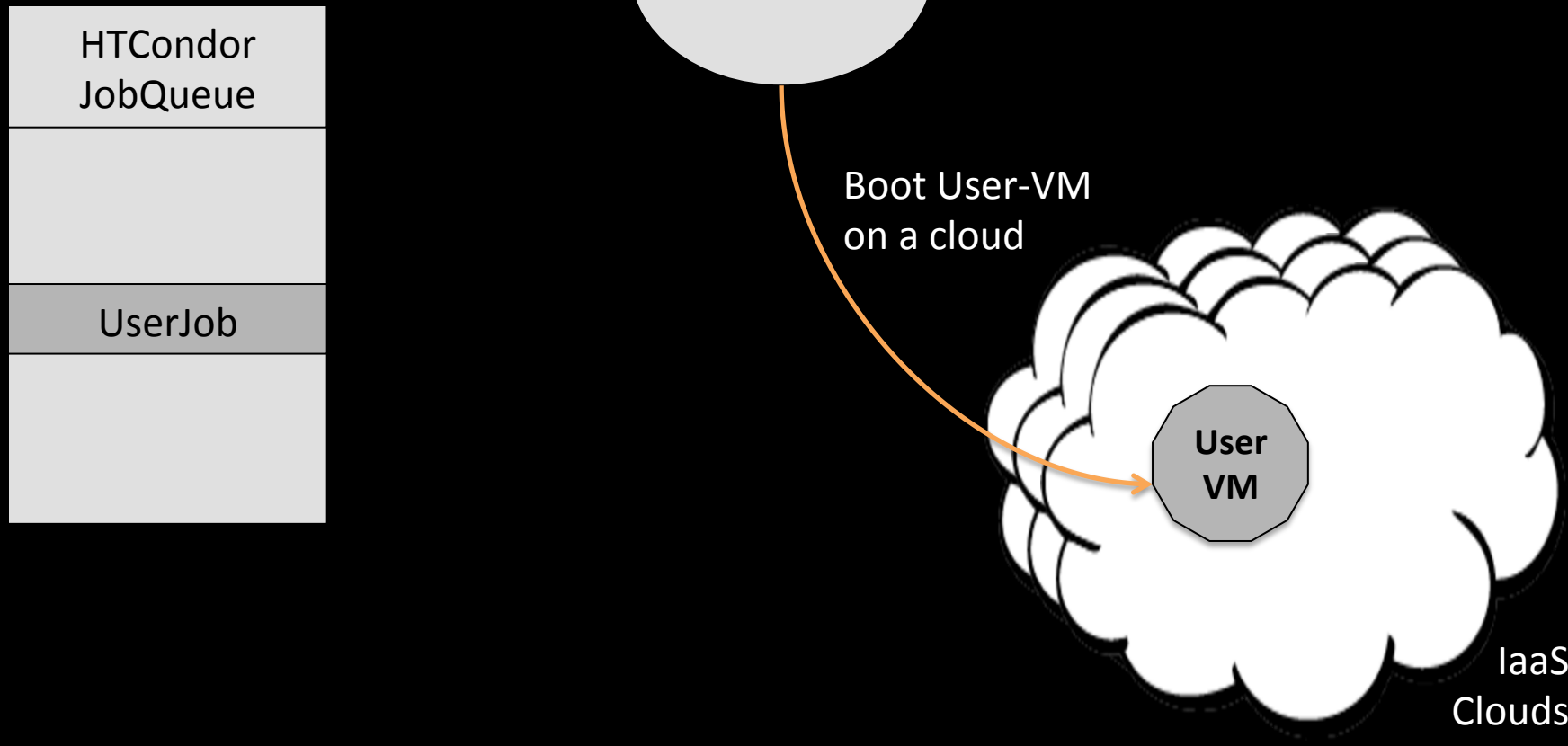
What are some of the challenges?

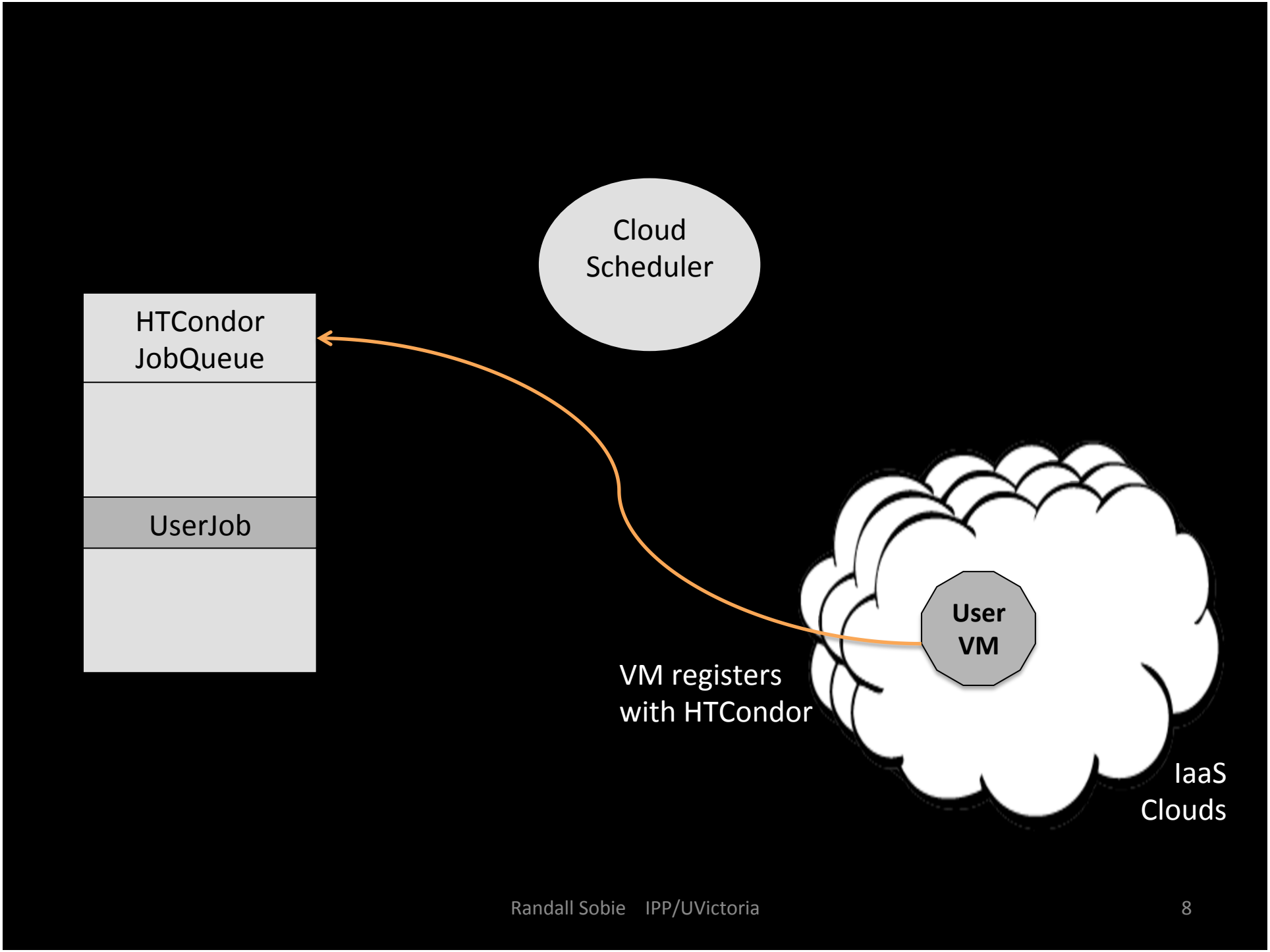
Distributed cloud computing system for ATLAS

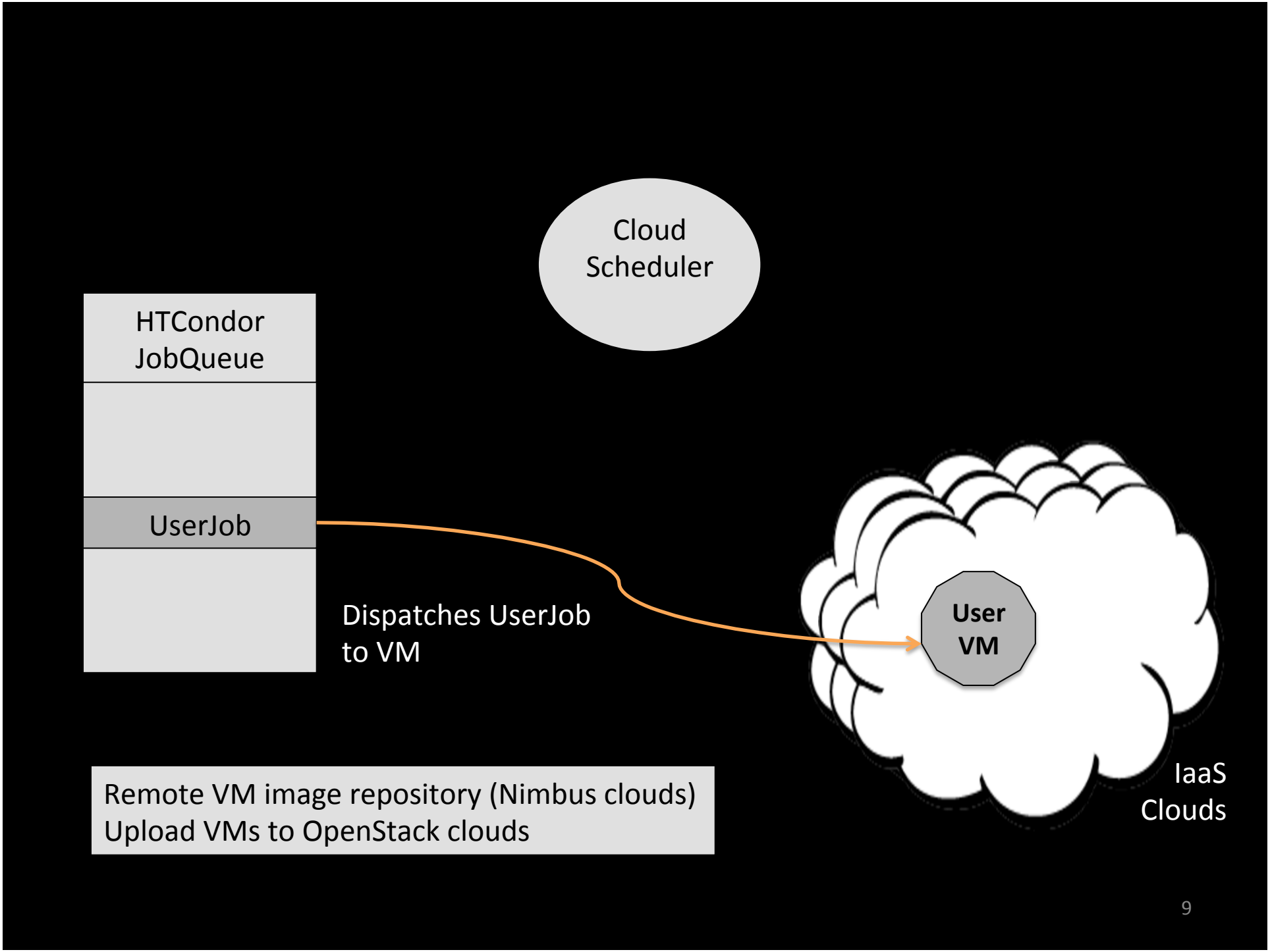
Production system for ATLAS experiment

Integrated into the WLCG

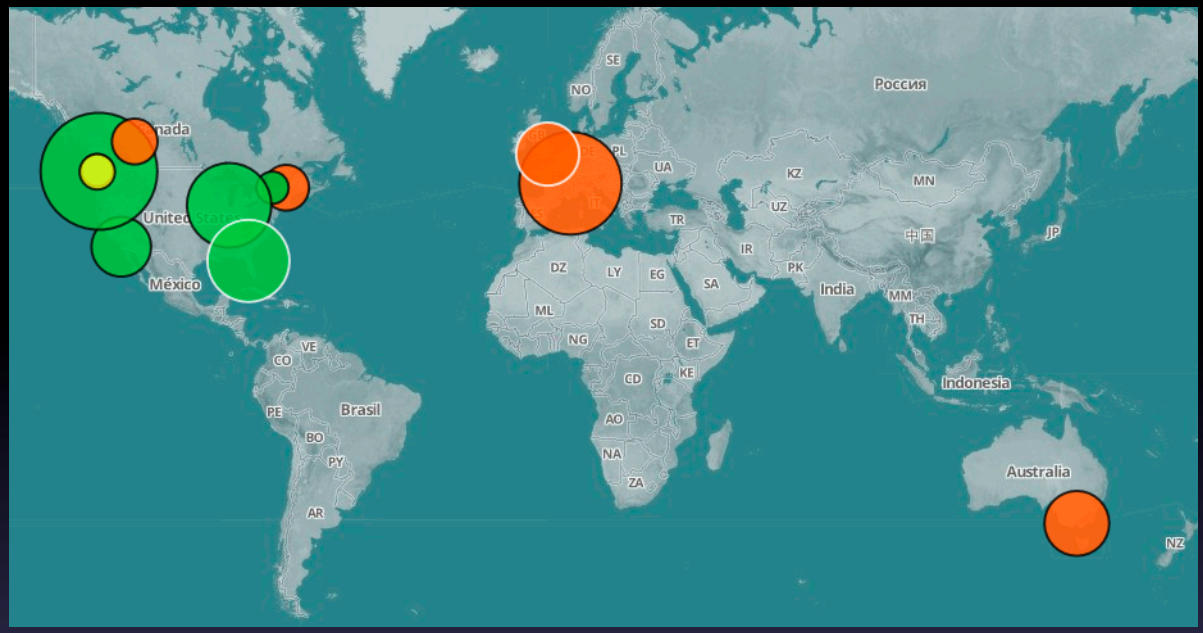

Cloud
Scheduler
HTCondor
JobQueue
Discover User-Job
UserJob
User
VM
IaaS
Clouds


HTCondor
JobQueue
UserJob
Cloud
Scheduler
Boot User-VM
on a cloud
User
VM
IaaS
Clouds


Cloud
Scheduler
HTCondor
JobQueue
UserJob
User
VM
VM registers
with HTCondor
IaaS
Clouds


Cloud
Scheduler
HTCondor
JobQueue
UserJob
Dispatches UserJob
to VM
User
VM
IaaS
Clouds
Remote VM image repository (Nimbus clouds)
Upload VMs to OpenStack clouds

Current clouds



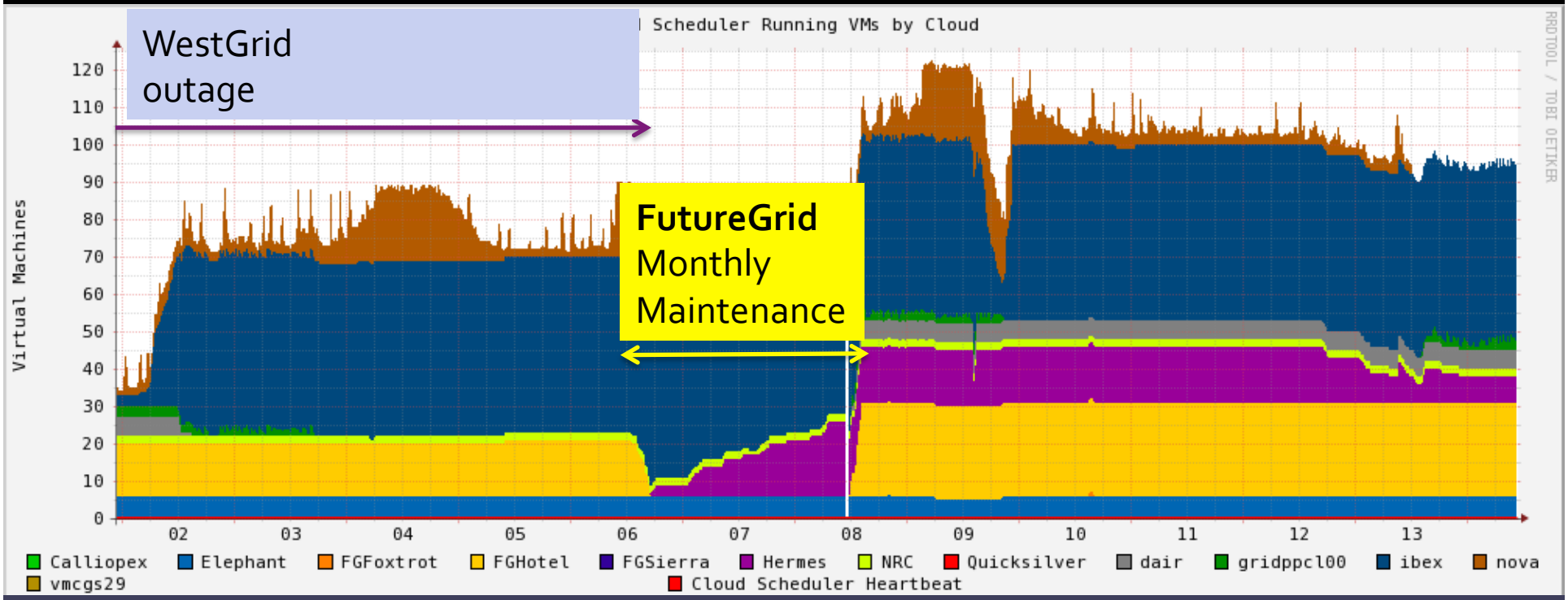
Nimbus

Victoria(3)
Ottawa
FutureGrid Chicago
FutureGrid SanDiego
FutureGrid Florida

OpenStack

Melbourne-NECTAR
CERN-Ibex
CANARIE-West
CANARIE-East
Imperial College-GridPP

Cloud evaporation and condensation



Number of 8-core VMs in May 2013

VMs in 2013

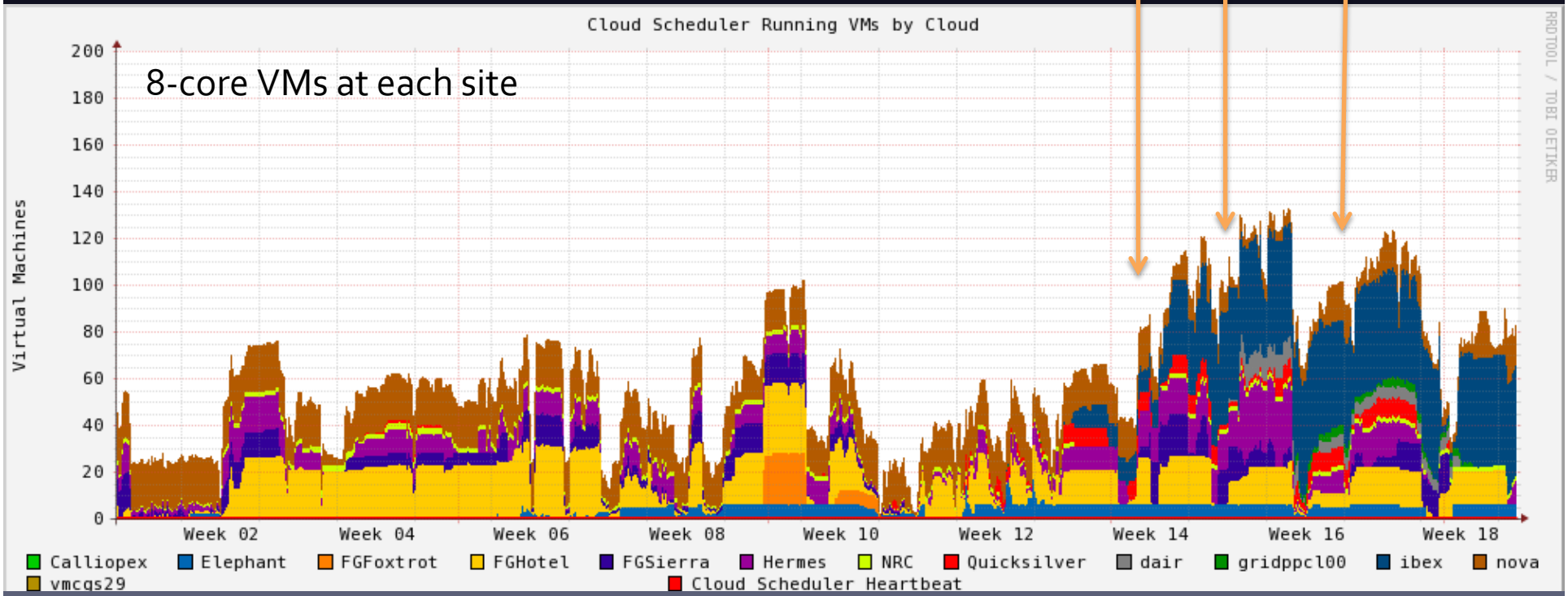
OpenStack Clouds

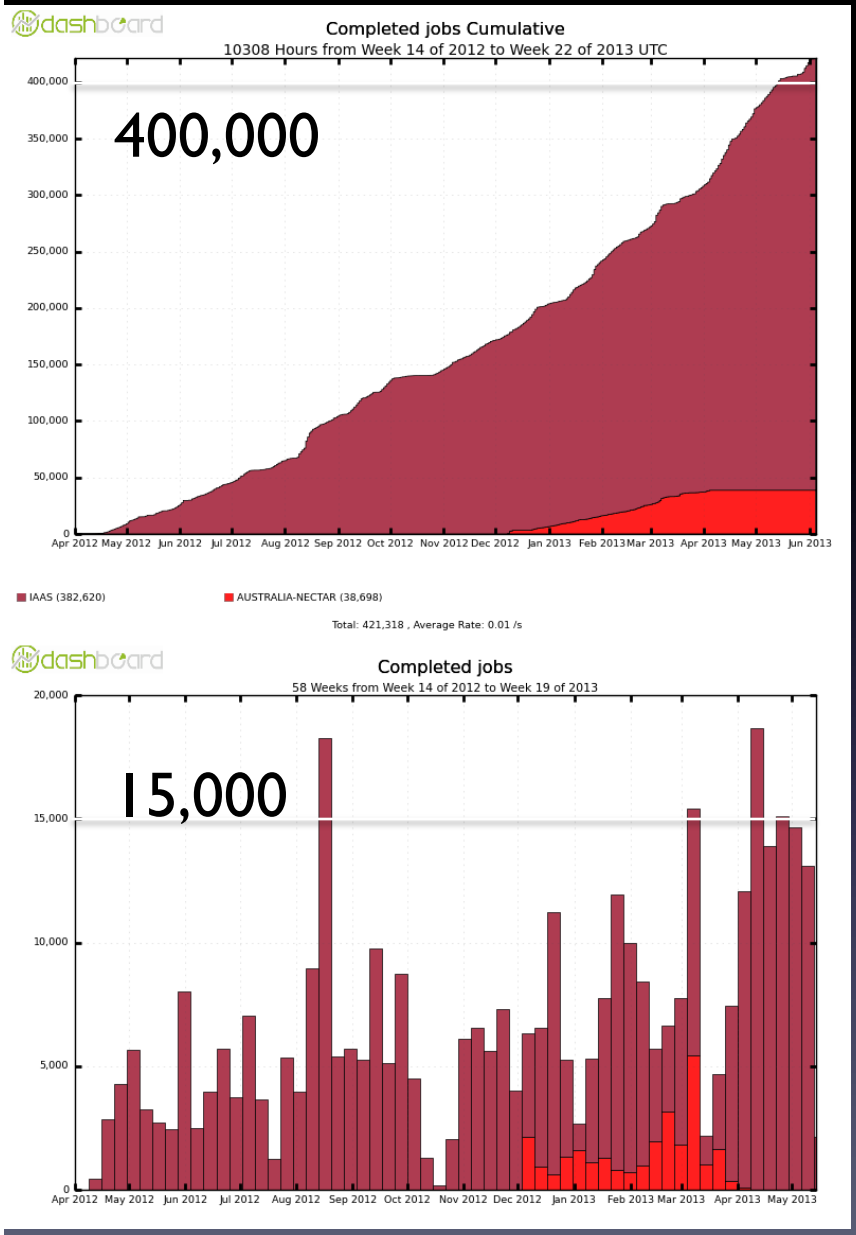
Nimbus Clouds (and NECTAR)

CERN IBex

CANARIE

Imperial College





Fully integrated as an ATLAS Grid site

(grid operations, monitoring, …)
April 2012

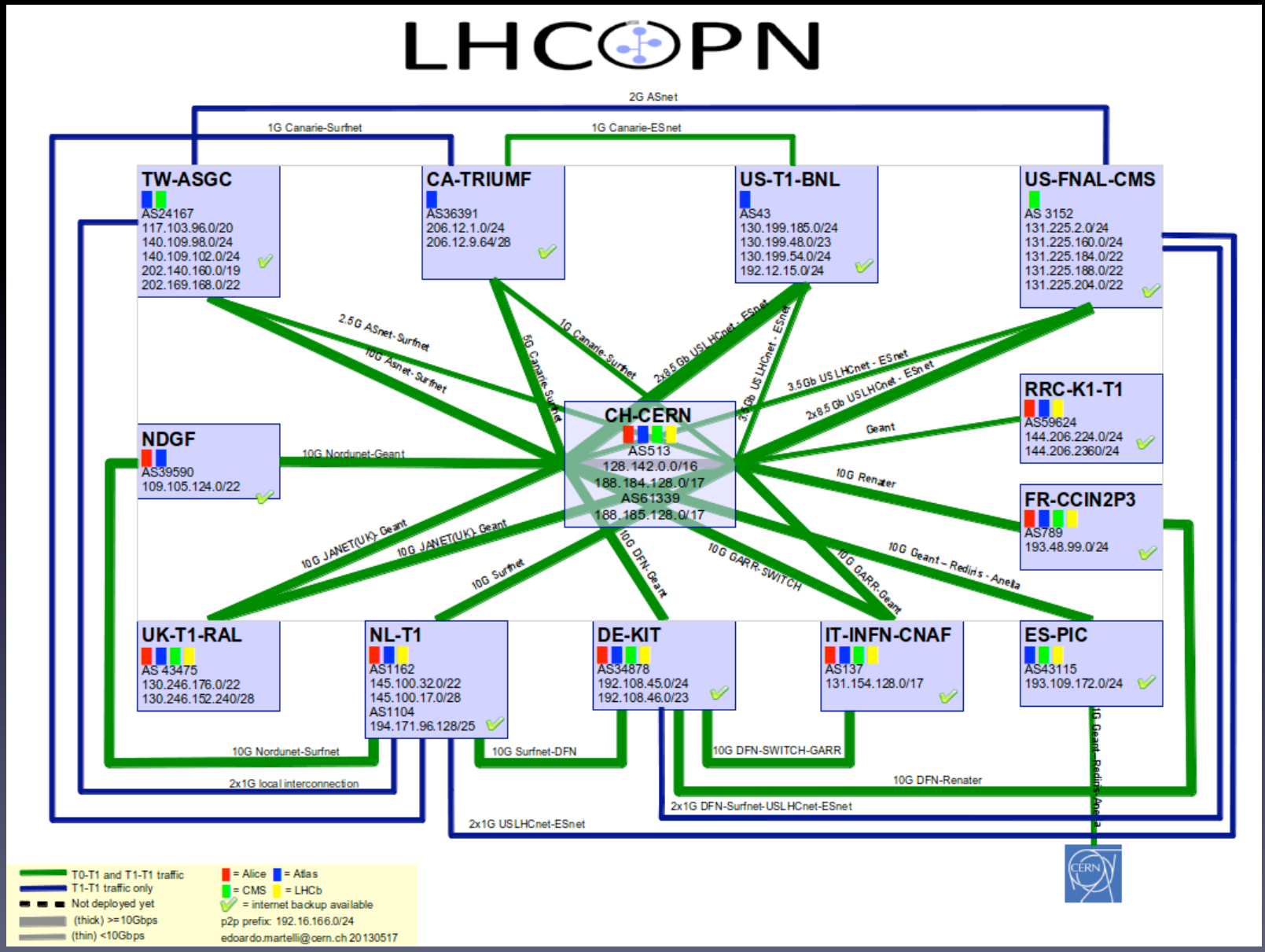
Integrated number of jobs (380,000)

Weekly jobs (12,500 in 2013)

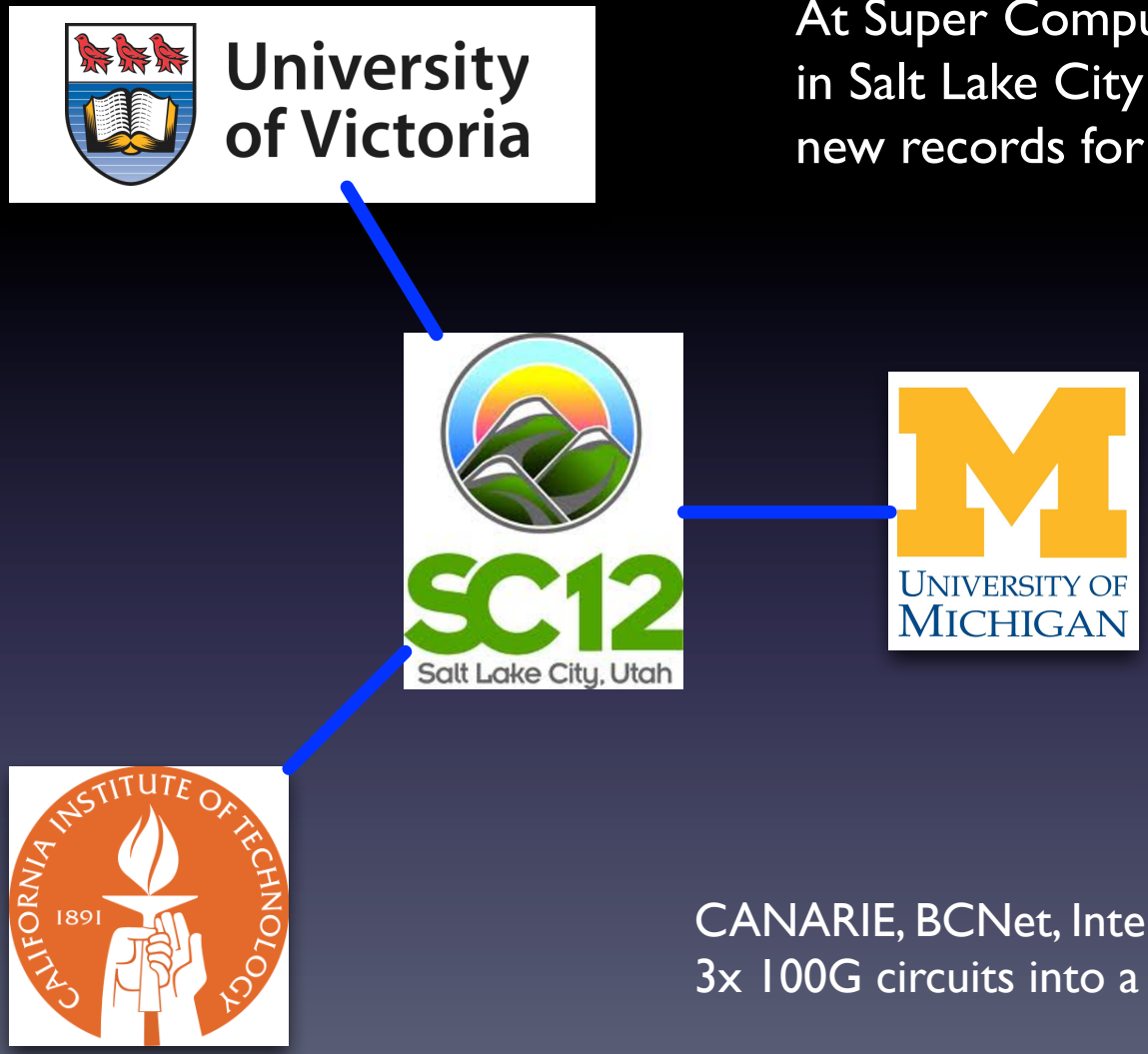
Peak over 1000 simultaneous jobs
April 2013

Moving on to Networking

HEP networks today



Next Generation 100G Networks

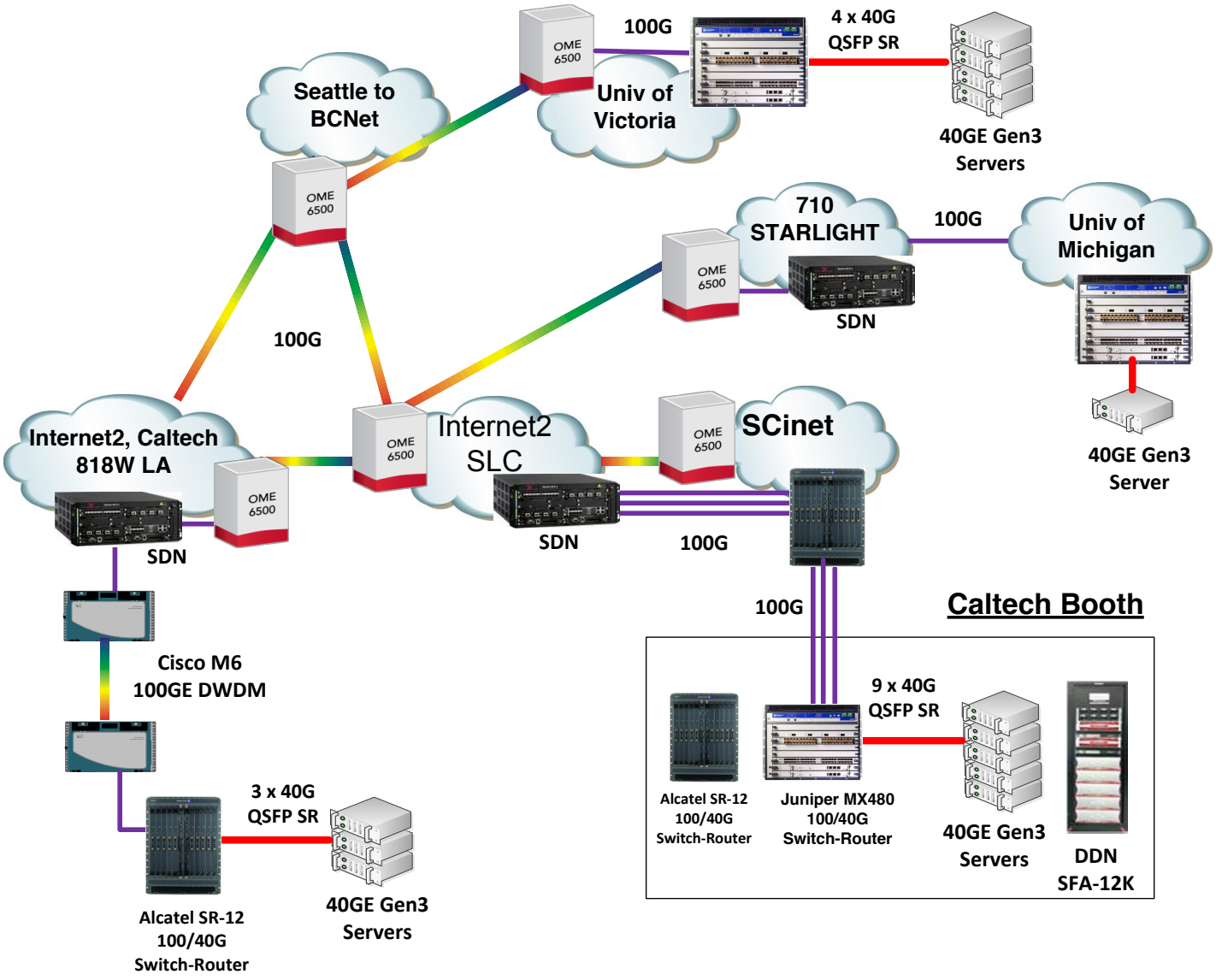


At Super Computing 2012 in November 2012 in Salt Lake City a we worked together to set new records for Wide Area Network transfer.

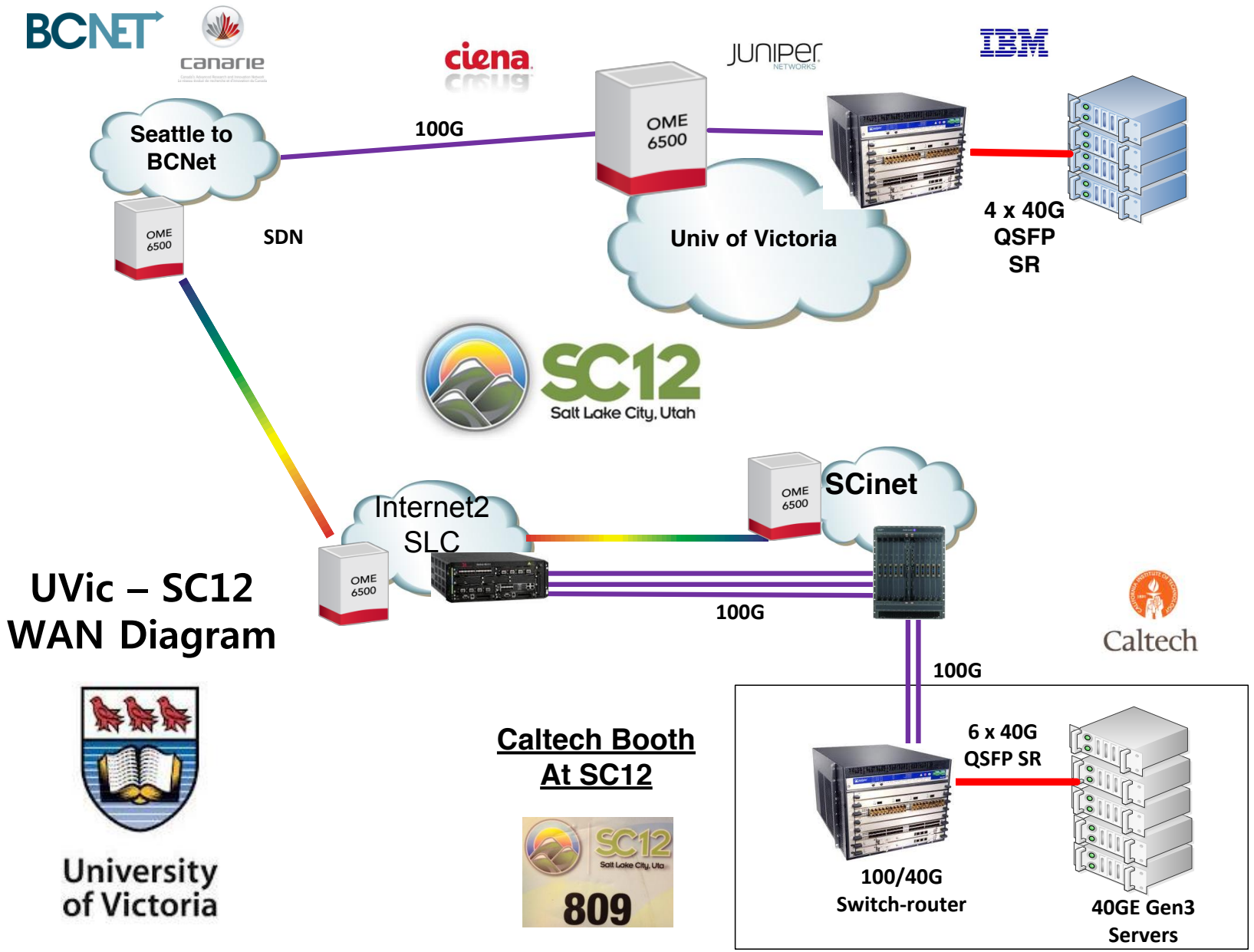
CANARIE, BCNet, Internet 2
3x 100G circuits into a single conference booth

SC12 WAN

Caltech – Victoria – Michigan
Efficient LHC Data Distribution across 100GE Networks



Victoria to Salt Lake City

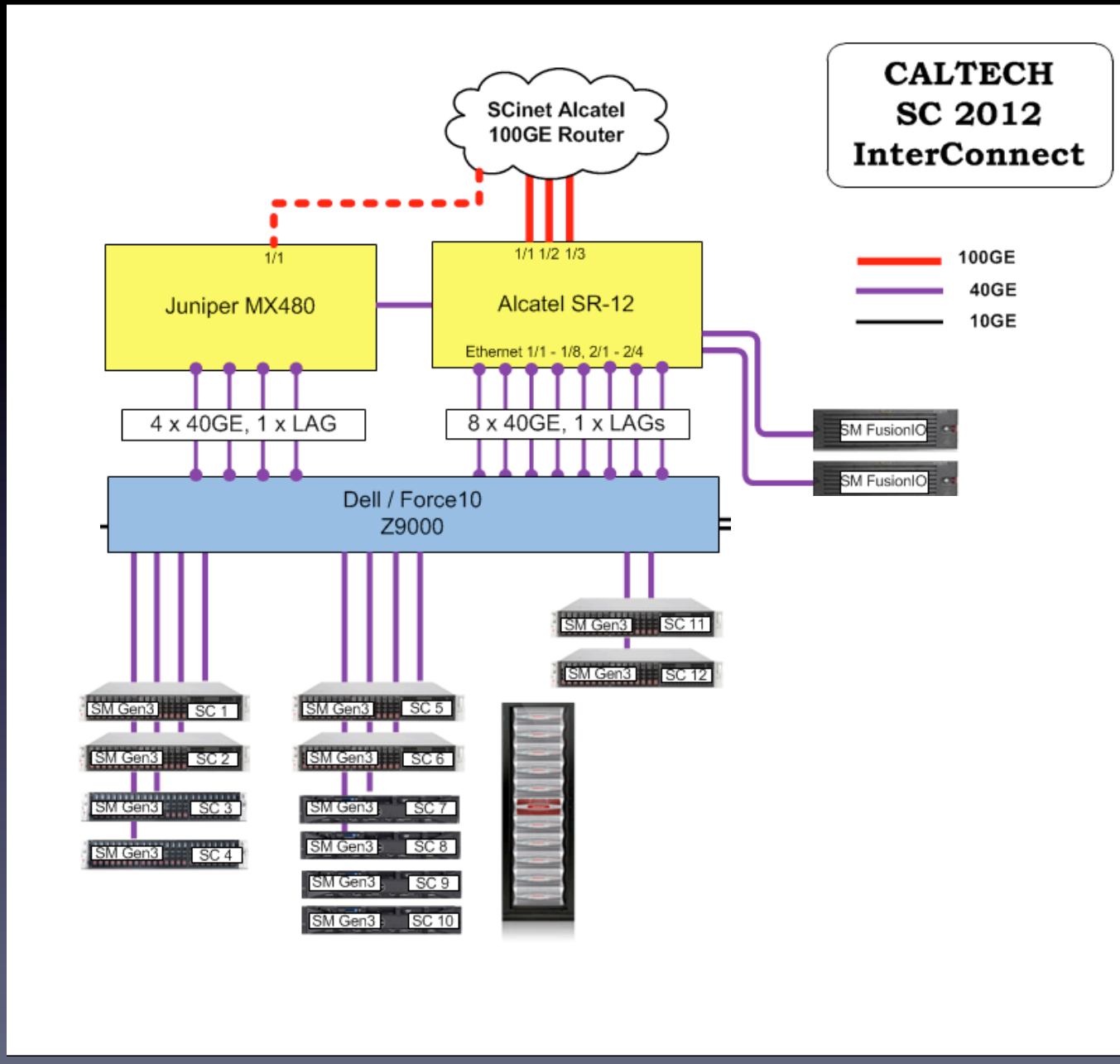


IBM x3650 M4

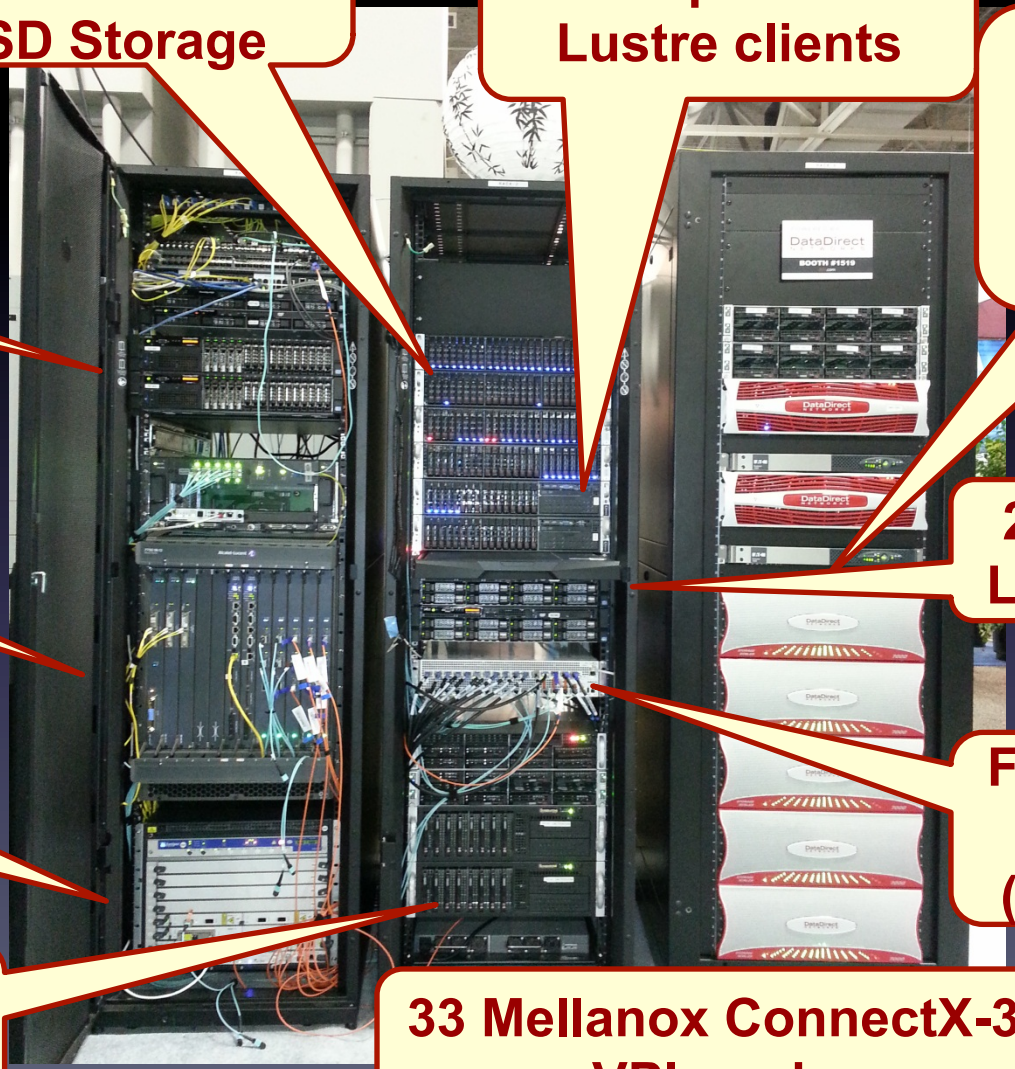
Juniper MX480

Ciena OME 6500


DataDirect
BOOTH #1519


CALTECH
SC 2012
InterConnect
SCinet Alcatel
100GE Router
1/1
Juniper MX480
1/1 1/2 1/3
Alcatel SR-12
Ethernet 1/1 - 1/8, 2/1 - 2/4
100GE
40GE
10GE
4 x 40GE, 1 x LAG
8 x 40GE, 1 x LAGs
SM FusionIO
SM FusionIO
Dell / Force10
Z9000
SM Gen3 SC 1
SM Gen3 SC 2
SM Gen3 SC 3
SM Gen3 SC 4
SM Gen3 SC 5
SM Gen3 SC 6
SM Gen3 SC 7
SM Gen3 SC 8
SM Gen3 SC 9
SM Gen3 SC 10
SM Gen3 SC 11
SM Gen3 SC 12

SC Show Floor 40G and 100G equipment



4x SuperMicro Servers SSD Storage

2x SuperMicro Lustre clients

Data Direct Networks Storage 120 TB

2x Dell R720 Lustre clients

Alcatel-Lucent 3x100GE 15x40GE

2x Dell R520 Lustre clients

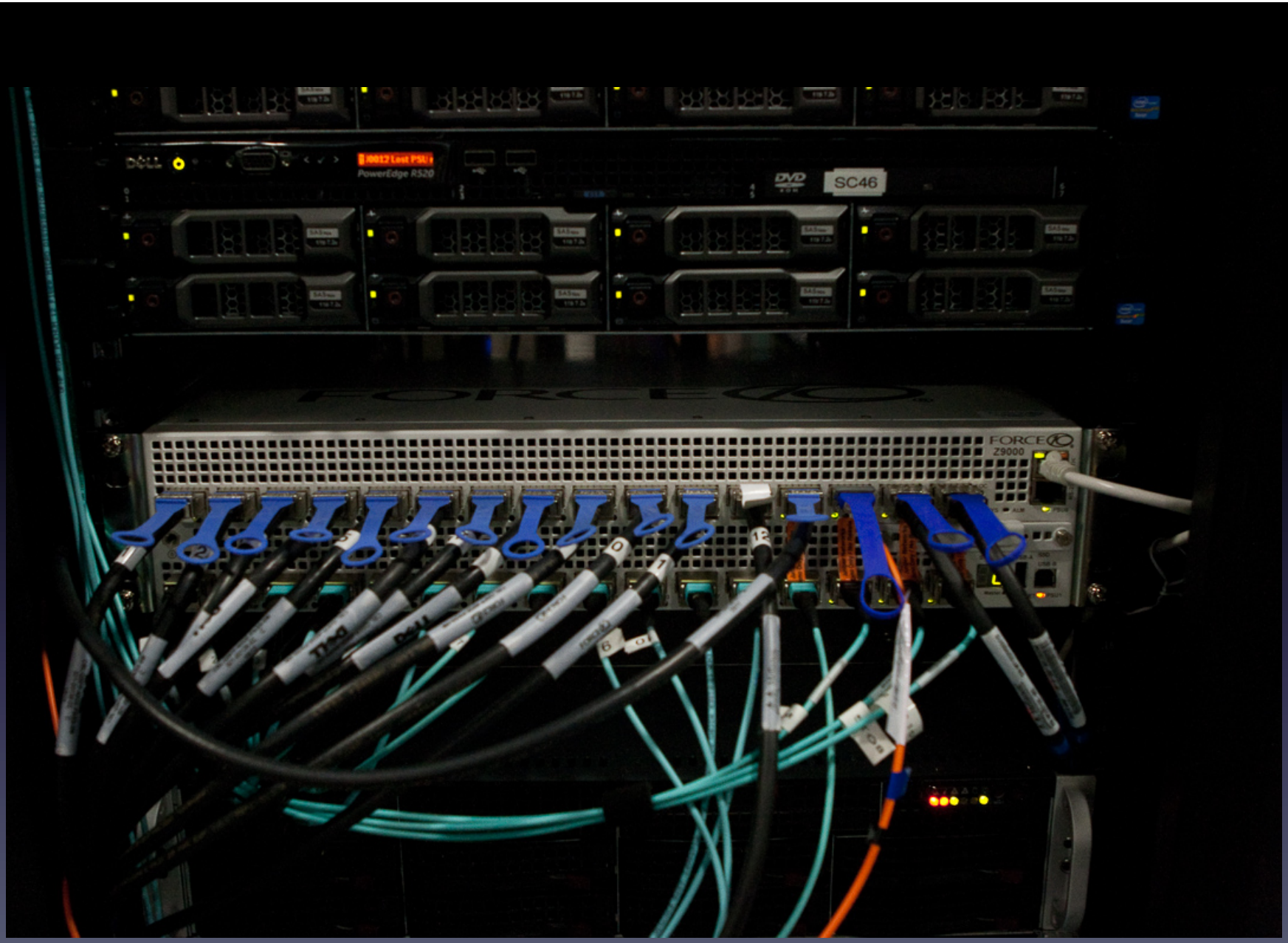
Juniper MX480 1x100GE 4x40GE

Force10/Dell Z9000 (40GE x 32)

2x SuperMicro Servers FusionIO Storage

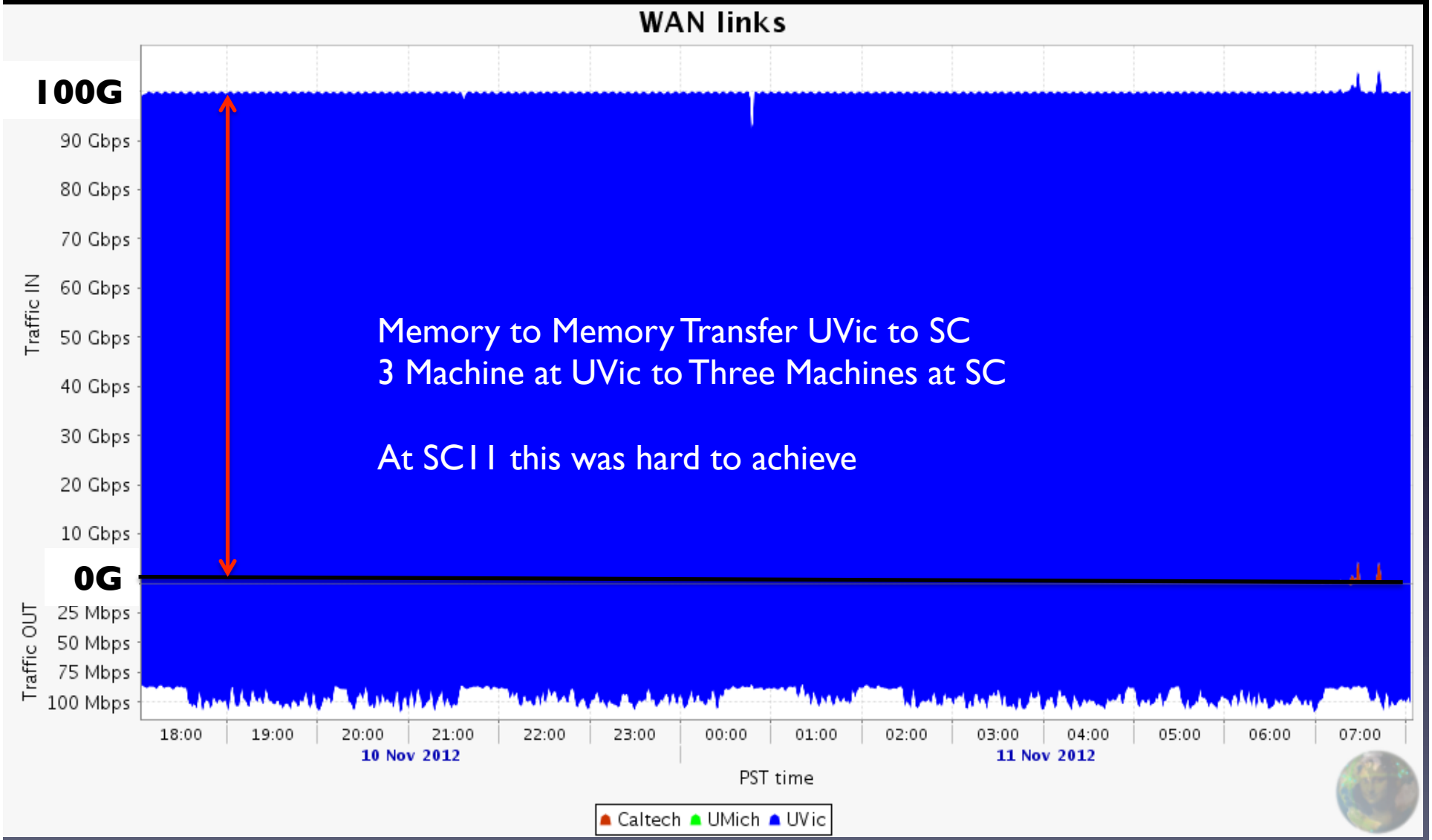
33 Mellanox ConnectX-3 VPI cards

Alcatel 3 x LR4 100G CFP

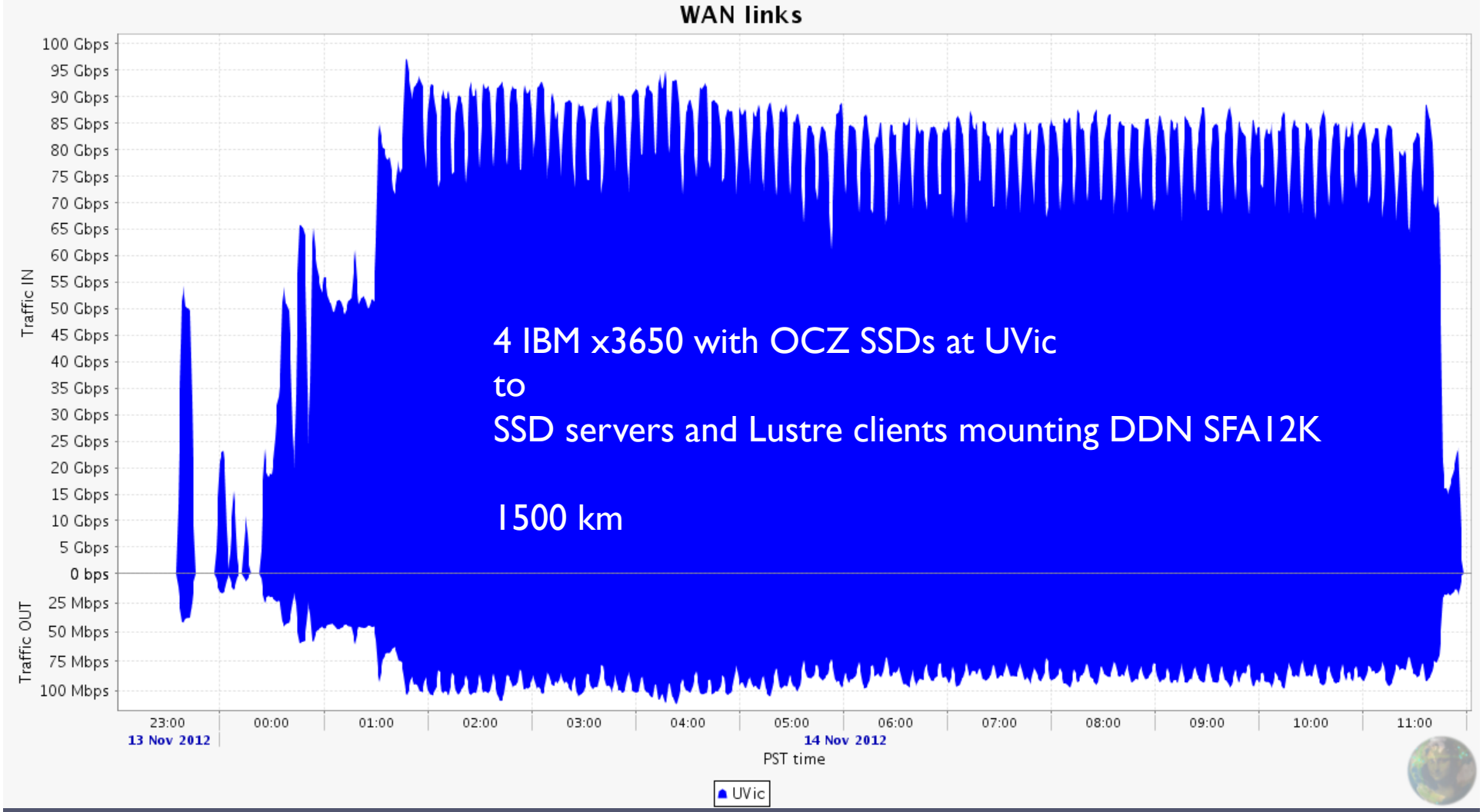


Force 10 z9000 32x40G Ethernet

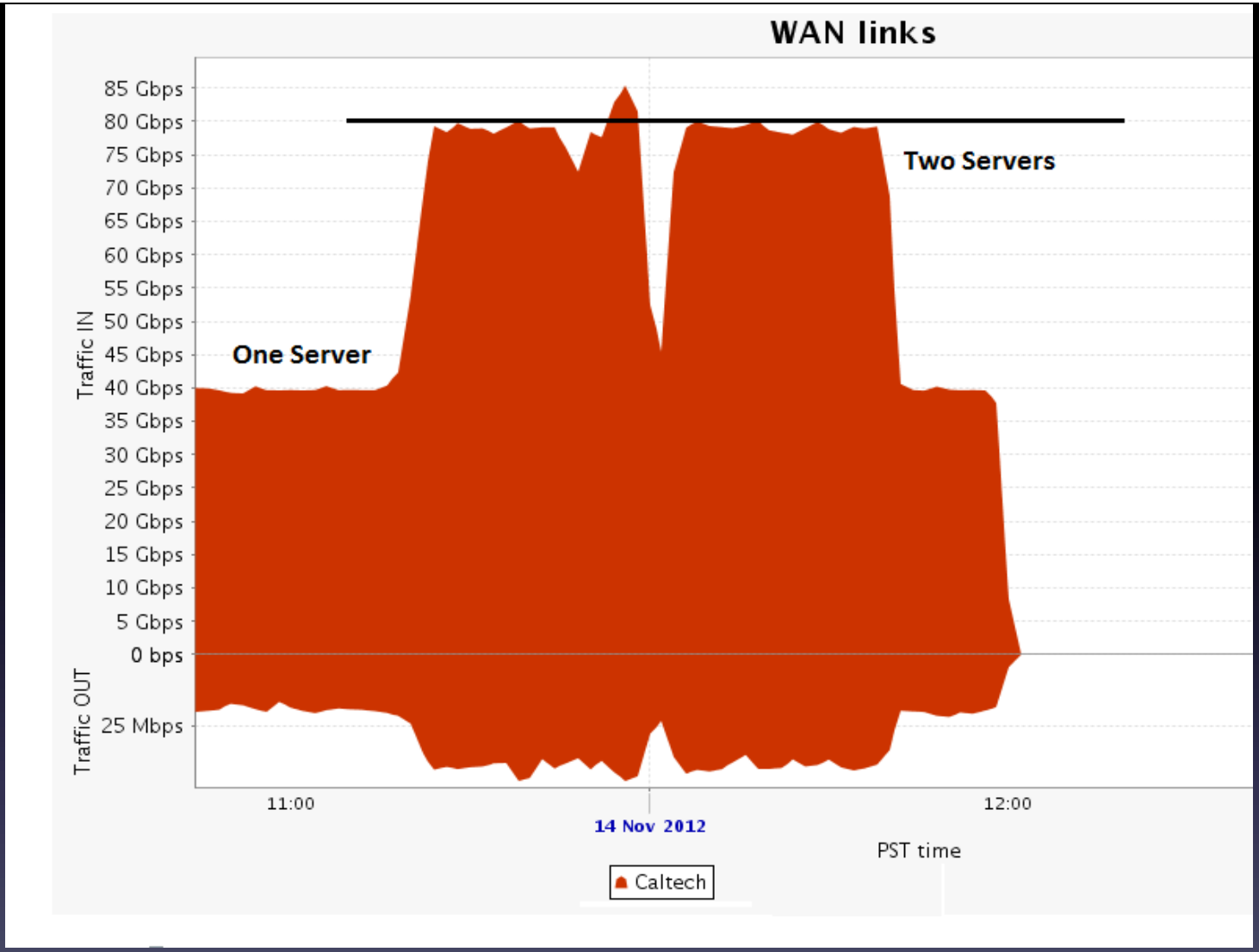
Day One: Memory to Memory



Disk to Disk



Peak 96 Gbps sustained 85 Gbps for 10 hours



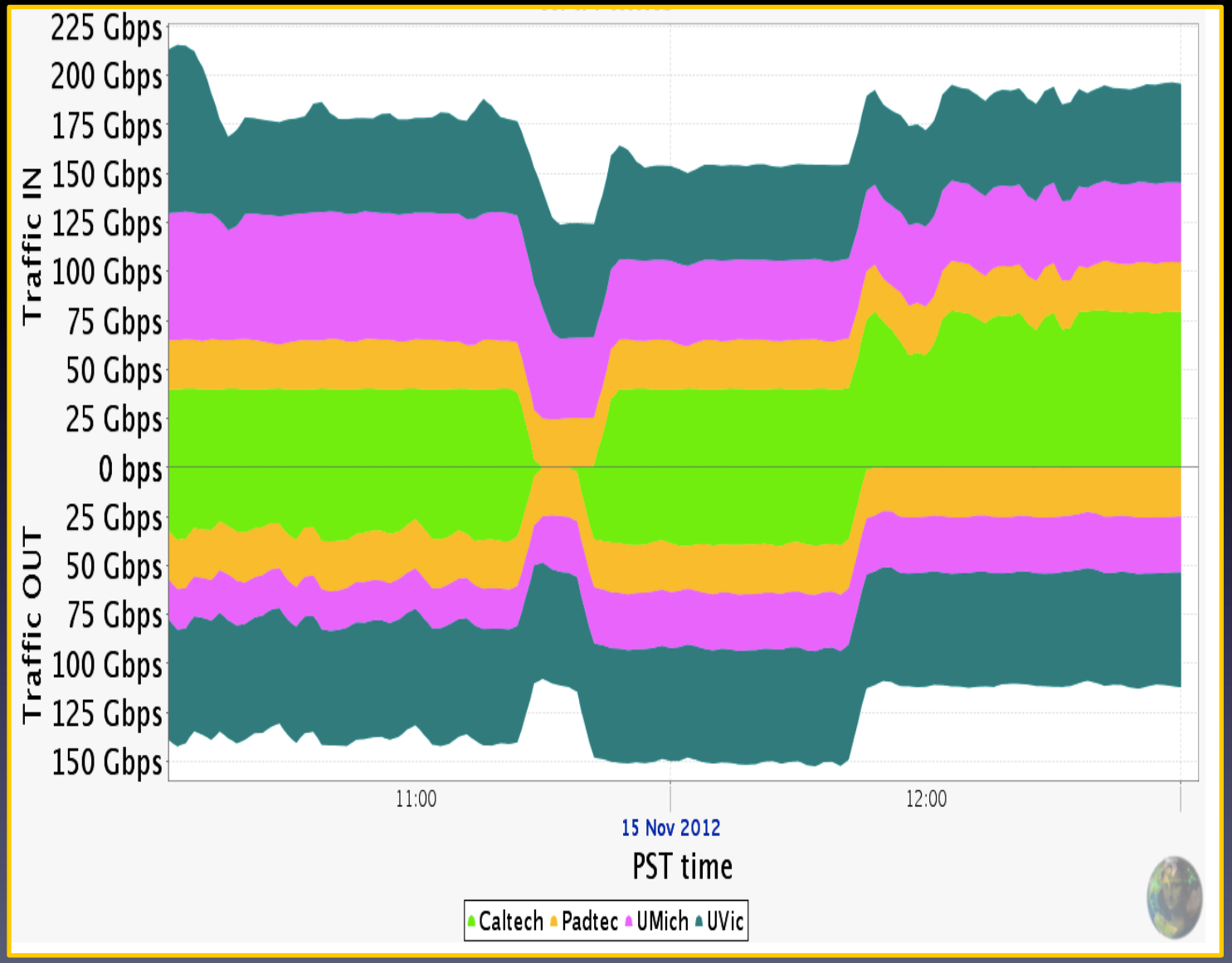
Remote Direct Memory Access over Ethernet

5% CPU Utilization

1 Server in the Booth and two Servers at Caltech

337 Gbps peak Memory to Memory

Rate equivalent to 3PB per day



Successes

IaaS Production system

- ATLAS IaaS distributed cloud in production for 14 month
- CANFAR (astronomy) using the same technology
- Federated system 10+ clouds over 3 continents, HEP and non-HEP sites
- 1000 simultaneous jobs
- Dynamic system that handles changing availability

100G networking

- 96 Gbps disk to disk using 'modest' set of hardware over 1500 km
- 337 Gbps memory to memory using 3 sites
- Promising demonstration of RDMA over Ethernet.
- Possible to efficiently use 100G networks at the end site

Summary

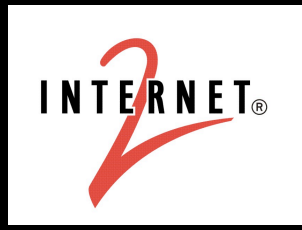
Clouds are increasingly available to the research community.

Distributed cloud computing is a viable solution for scientific computing

No observed limits to the system scalability in low I/O cases.

We expect 100G networks to remove barriers to high I/O applications on clouds.

Industrial Partners



http://heprc.phys.uvic.ca
igable@uvic.ca
@igable

Cloud Scheduler:
https://github.com/hep-gc/cloud-scheduler