



# An Operational Perspective on a Hybrid and Heterogeneous Cray XC50 System

CUG 2017

Sadaf Alam, Nicola bianchi, Nicholas Cardo, Matteo Chesi, Miguel Gila, Stefano Gorini Mark Klein, Colin McMurtrie, Marco Passerini, Carmelo Ponti, Fabio Verzelloni, CSCS

May 10, 2017



*This presentation will touch on a lot of areas. Please feel free to stop and talk to us, email us, or call us if you would like to chat more in detail.*

# Agenda



- Piz Daint Upgrade
- Piz Dom (TDS)
- Innovative Features
  - Public IPs
  - Enhanced GPU Monitoring
  - Slurm Configuration
- Advanced Services
  - Compute Acceleration
  - Visualization and Data Analytics
  - Data Transfer Service
  - Container
- Cray Sonexion 3000

## Piz Daint Upgrade

---



# Basic Concept

- Seems simple...
  - Replace Haswell with Broadwell
  - Replace Sandybridge with Haswell
  - Replace K20X with P100
  - Change XC30 cabinets for XC50
  - Add Sonexion 3000 scratch
- Did I mention...
  - Merge Piz Dora into Piz Daint
- Let's not forget...
  - ~~▪ Do it all in under 3 months.~~
- By the way...
  - SLES goes from 11 to 12
  - xtopview replaced with Ansible
  - XC30 to XC50 (bigger cabinet)
  - XC40 (Dora) relocated
  - Retain small scale production operations
- In case I forgot...
  - Do it all in 2 months

## Haswell to Broadwell

- Replace Intel(R) Xeon(R) CPU E5-2690 v3 @ 2.60GHz (12 cores, dual socket) with Intel(R) Xeon(R) CPU E5-2695 v4 @ 2.10GHz (18 cores, dual socket)
- Straight forward due to socket compatibility
- Completed ahead of the main upgrade.

## XC30 -> XC50 Upgrade

- Remove all XC30 cabinets
- Install new XC50 cabinets
- Simple
  - Shutdown
  - Disconnect cabinets
  - Physically remove XC30 cabinets (*and cables for everything*)
  - Physically install XC50 cabinets
  - Re-cable (*1632 just for Aries*)
  - Merge with Piz Dora
  - Boot
- While we are at it...
  - Add a second SCRATCH filesystem on Sonexion 3000

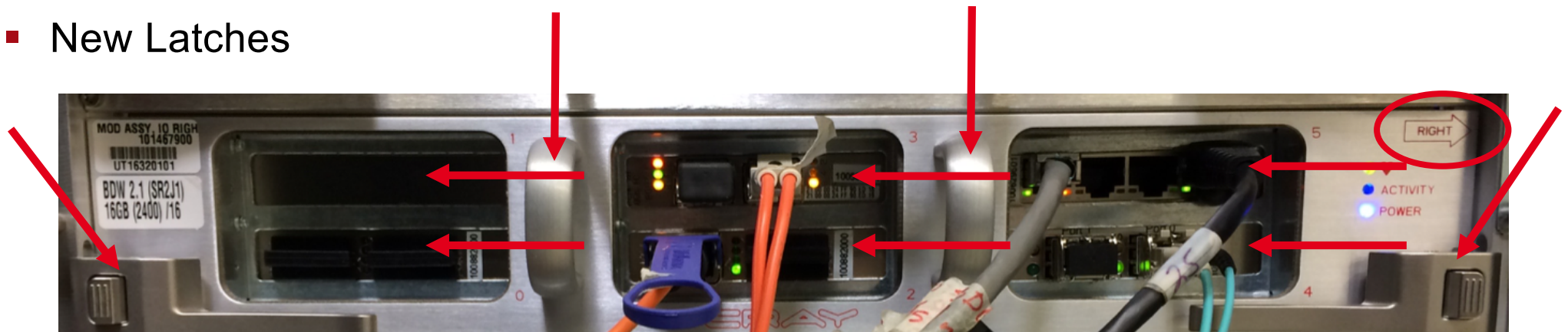
# XC50 Compute Blade



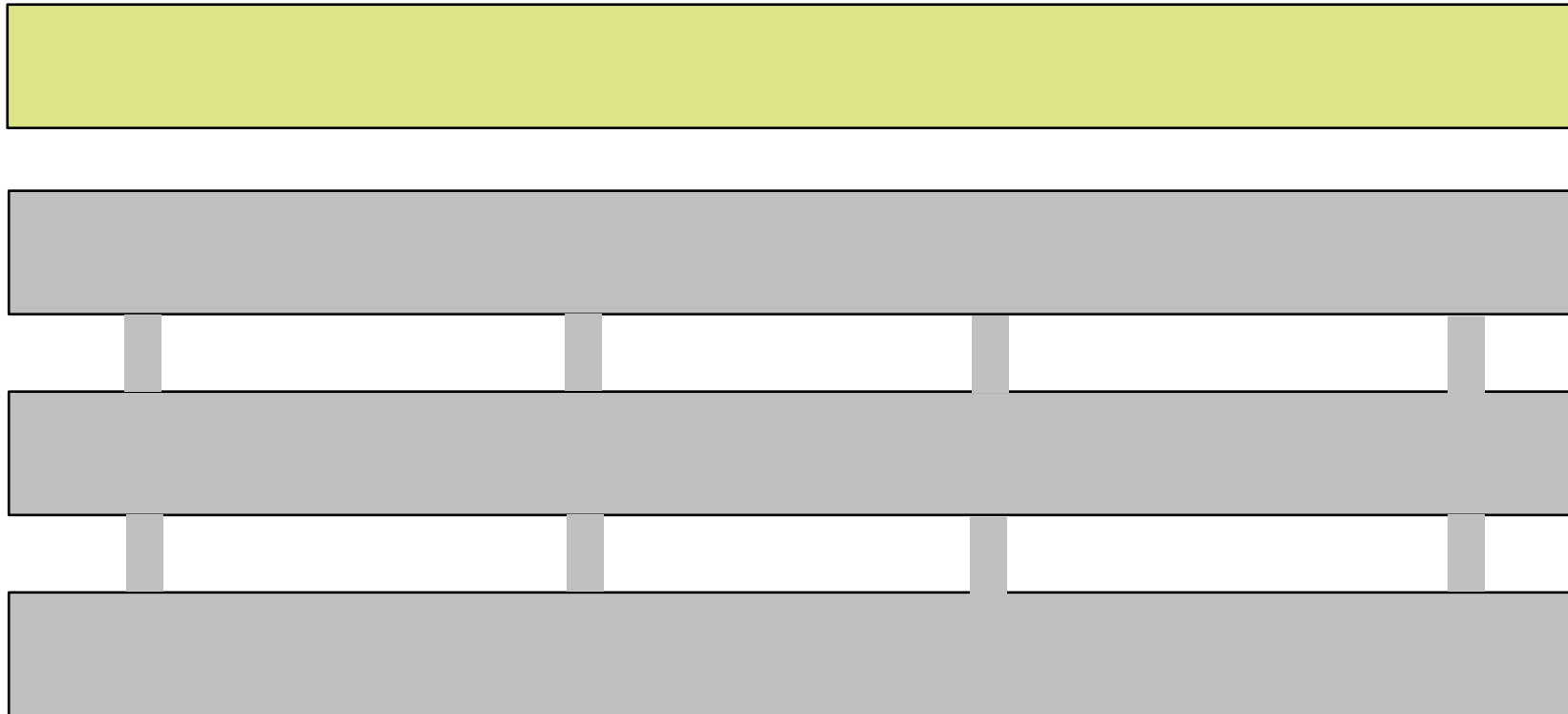


## New XC 50 SIO Module

- 4 Nodes
  - 2 nodes x 2 slots
  - 2 nodes x 1 slot
  - 64 GB
  - Intel(R) Xeon(R) CPU E5-2695 v4 @ 2.10GHz
- Left AND Right Modules 😊, but both are required ☹️
- Handles, will explain later...
- New Latches



# Upgrade to Broadwell



# Remove XC30



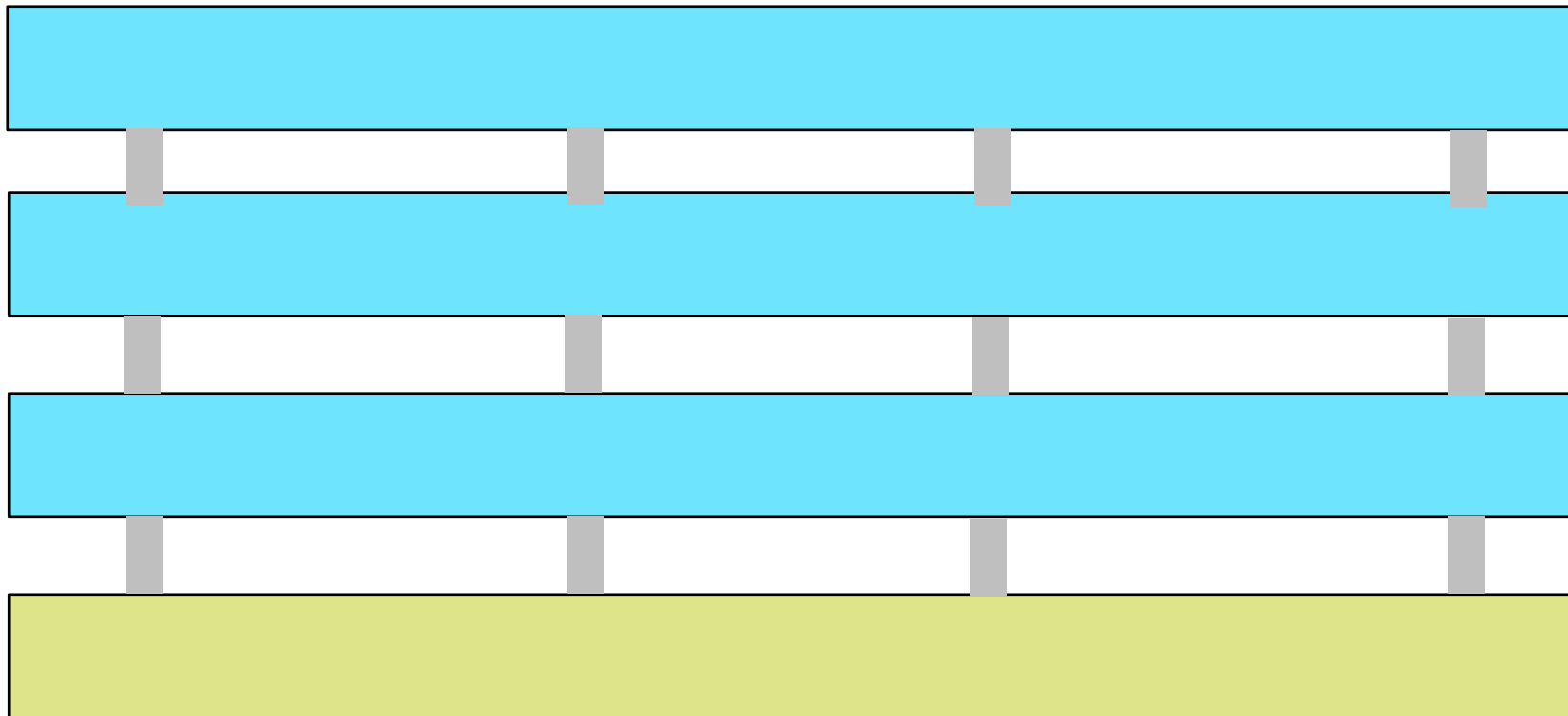
# Relocate XC40 Cabinets



## Bring in XC50 Cabinets



# Cable as a Single System

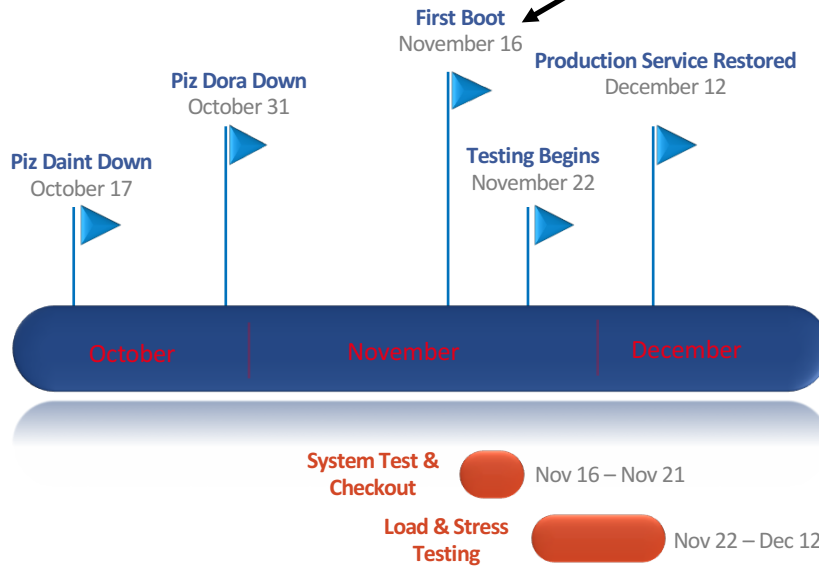


# Timeline



30 days!

Thanks to the hard work by a lot of people!



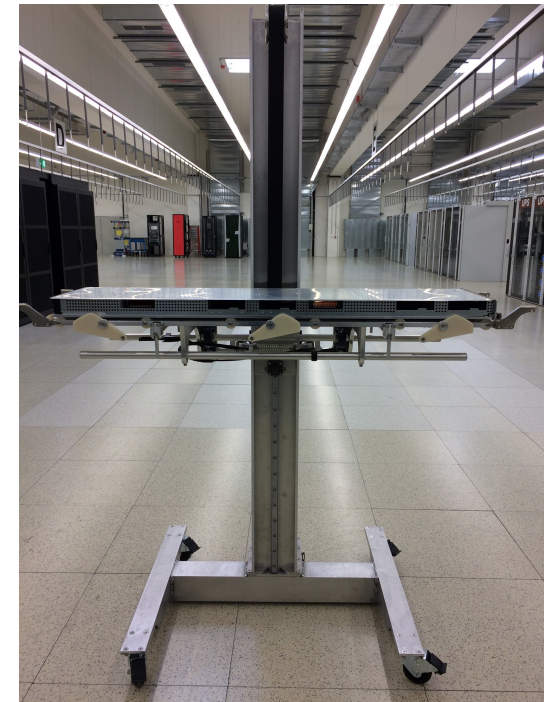
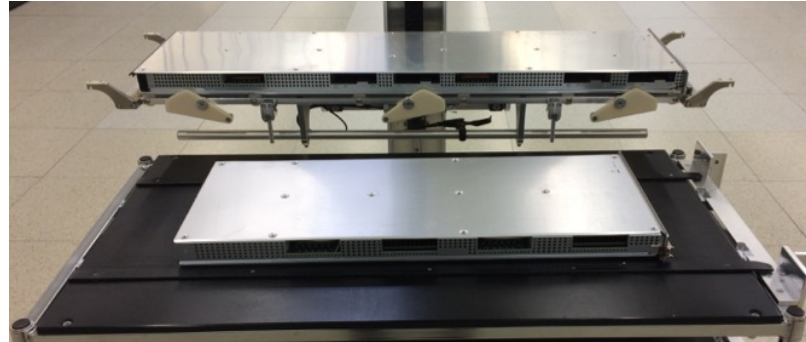
*Piz Daint Users Service Interruption: 56 days*  
*Piz Dora Users Service Interruption: 42 days*

**Completed in under 2 months!**



# Fun Challenges

*Bigger Modules*



*Requires Special Handling*

*Bigger Cabinets*







**CSCS**

Centro Svizzero di Calcolo Scientifico  
Swiss National Supercomputing Centre

**ETH** zürich

**Piz Dom**

---



# Test Environment



Single Cabinet Single Chassis XC50  
Single Cabinet Single Chassis XC40  
Configured as 2 Rows



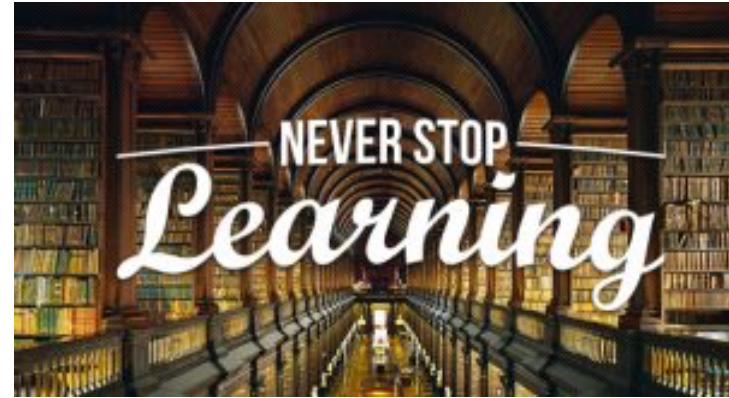
**CSCS**

Centro Svizzero di Calcolo Scientifico  
Swiss National Supercomputing Centre

**ETH** zürich

## Innovative Features

---





Strive for perfection in everything.  
Take the best that exists and make it  
better. If it doesn't exist, create it.  
Accept nothing nearly right or good  
enough

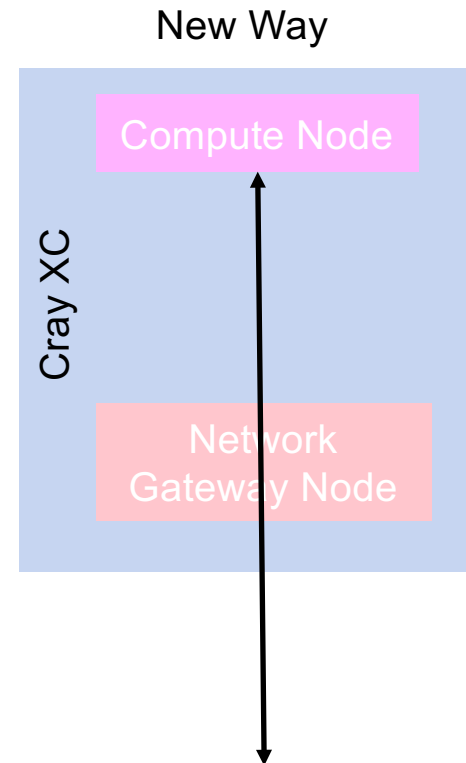
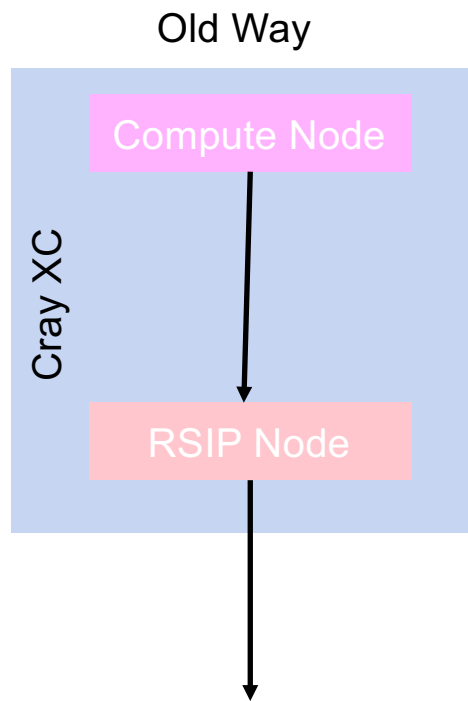
— *Henry Royce* —

AZ QUOTES

## Enhanced GPU Monitoring

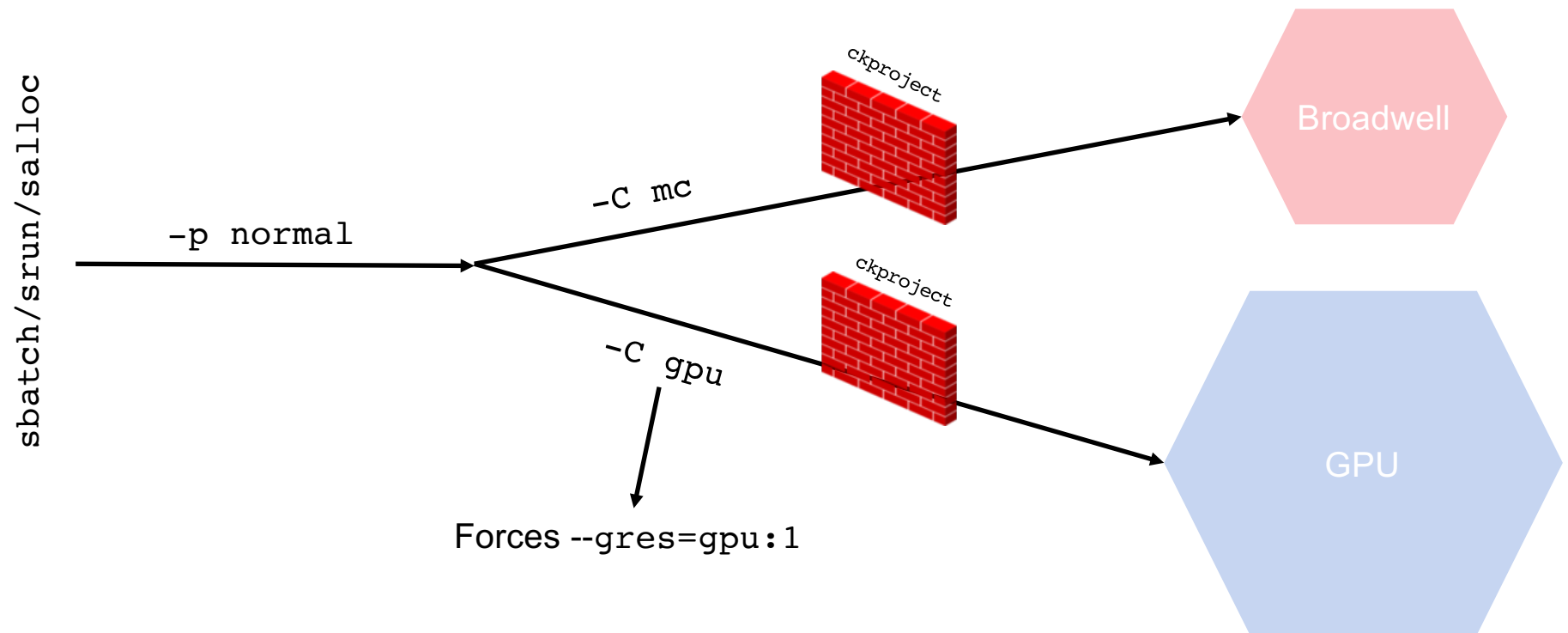
- **The Problem:**
  - Very limited GPU checking has resulted in nodes with problematic GPUs being allocated to more batch jobs.
  - Relies on users reporting problems instead of detecting problems.
- **The Solution:**
  - GPUs are checked for a number of error conditions as part of job startup and job exit.
  - A suspect GPU will result in that node automatically being taken out of service and the error reported.
  - Detects total failures as well as suspect failure modes. Many cases nodes just need a reboot to cleanup. Evaluates XID errors, GPU memory, CUDA driver version, GPU Inforom and VBIOS, hung processes in the GPU, general GPU health, PCI-e link width)
  - Checks can also be run manually/interactively.
- **The Results:**
  - Extremely Successful!
  - Expanded to include node issues.
  - Sample Errors:
    - `cksys system=daint,node=nid02455,GPU S/N=0323616005254,jobid=1412154,user=(null),test=ckgpuXID,msg=reboot/retest: XIDs 45 48 63 64`
    - `cksys system=daint,node=nid05622,GPU S/N=0323216058722,jobid=1401812,user=(null),test=ckgpuhealth,msg=reboot and retest`

# Public IP Routing



- Simplifies SLURM
- Uses standard networking
- Easy application steering
- Easy license management
- Possibility Enabler

# Native Slurm Configuration





**CSCS**

Centro Svizzero di Calcolo Scientifico  
Swiss National Supercomputing Centre

**ETH** zürich

# Advanced Services

---



## Compute Acceleration

Feature	Telsa K20X	Tesla P100
Share Memorys	12	56
Base Clock	732 MHz	1328 MHz
GPU Boost Clock		1480 MHz
Double Precision Perf.	1.31 TF	4.7 TF (not DGEMM)
Single Precision Perf.	3.95 TF	9.3 TF
Half Precision Perf.		18.7 TF
PCIe X16 B/W	16 GB/s bi-dir PCI Gen 2.0	32 GB/s bi-dir PCI Gen 3.0
Memory Interface	384 bit GDDR5	4096 bit (CoWoS HBM2)
Memory Capacity	6 GB	16 GB
Memory Bandwidth	250 GB/s	732 GB/s

*Out with Intel(R) Xeon(R) CPU E5-2670 @ 2.60GHz/K20X*

*In with Intel(R) Xeon(R) CPU E5-2690 v3 @ 2.60GHz/P100*

# Visualization and Data Analysis Services


- Visualization
  - Available on all GPU enabled compute nodes.
  - Selectable by SLURM Job constraint. Prolog automatically performs the setup.
  - EGL enabled driver.
- Data Analysis
  - Deep Learning applications provide a GPU-accelerated library of primitives for deep neural networks.
  - NVIDIA CUDA Deep Neural Network (cuDNN) is a highly tuned implementation.
  - GPU accelerated frameworks have been accelerated using cuDNN: Caffe, CNTK, TensorFlow...

# Data Transfer Service

- Motivations:
  - Need to transfer data efficiently.
  - Desire to transfer data without wasting valuable compute nodes.
  - Capability to integrate into production workflows.
  - Utilize standard transfer methods, no custom software development.
  - Integrate with Slurm.
- Solution:
  - Created a small cluster designed for this purpose that is expandable.
  - Initial supported transfer services: GridFTP, Copy, and Move.
  - Can use Job Dependencies to automatically trigger transfer and production jobs.
- Results:
  - Extremely successful!
  - Need to train users on how to best use this service.



## Container Service

- Providing Shifter for production computing use.
  - Developed by  and enhanced by CSCS to support GPUs.
  - Initial customer is the physics community which required a Red Hat operating system.
- Why Containers on a Cray?
  - The use of containerized computing is growing.
  - Extends the possibilities for use to support customer use case restrictions.
  - Opens the door to possibilities.
  - Enables tight control over the runtime environment.
  - In some cases, improves performance.

28A: Shifter: Fast and consistent HPC workflows using containers, 15:00 Thursday

## Cray Sonexion 3000

---



*Product Specifications: <http://www.cray.com/sites/default/files/SonexionBrochure.pdf>*

## The Appliance...

- Scalable Storage Unit (SSU)
  - 2 Lustre Object Storage Servers (OSS)
  - 82 8TB drives
  - Published Product Peak: 11-16 GB/s
  - Published Product Sustained: 9-14 GB/s
- Expansion Storage Unit (ESU)
  - Connects to SSU
  - 82 8TB drives
- Our Configuration
  - 7 SSU + 7 ESU
  - Packaged with 2 MetaData Servers (MDS) and 1 Management Server (MGS)
  - Total: 9.18 PB raw, 6.2 PiB usable.
  - Measured Performance (IOR):
    - Sequential Reads: 64 GB/s
    - Sequential Writes: 81GB/s

## Acknowledgement

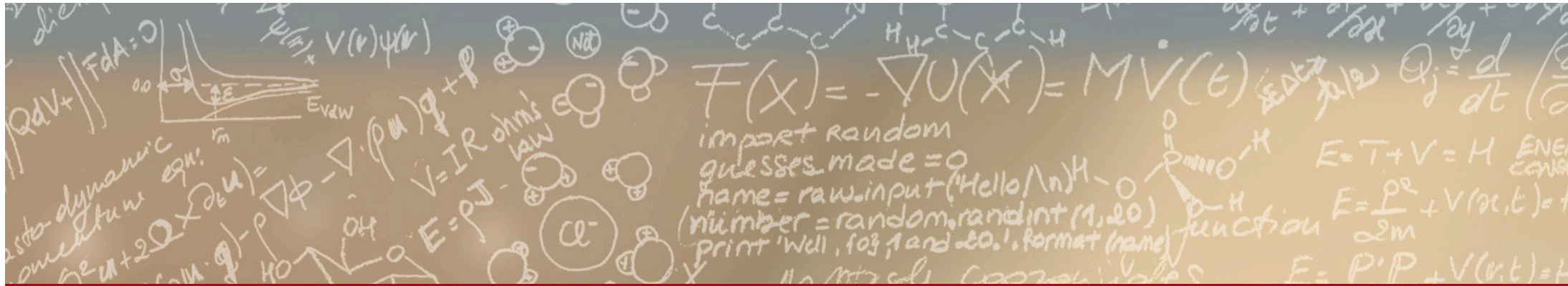
*The details provided in this presentation, and paper, are the results of the dedication and efforts of a lot of people.*

*A special thanks to all the Cray Engineers who worked collaboratively with CSCS to successfully accomplish the upgrade in a short time.*

*The research into providing the container service was supported by the Swiss National Science Foundation.*

*Thank  
you*





## Thank you for your attention.

Come and See Us!

BoF 20D: Bringing "Shifter" to the Broader Community, 17:10 Wednesday

27C: A regression framework for checking the health of large HPC systems, 13:00 Thursday

28A: Shifter: Fast and consistent HPC workflows using containers, 15:00 Thursday