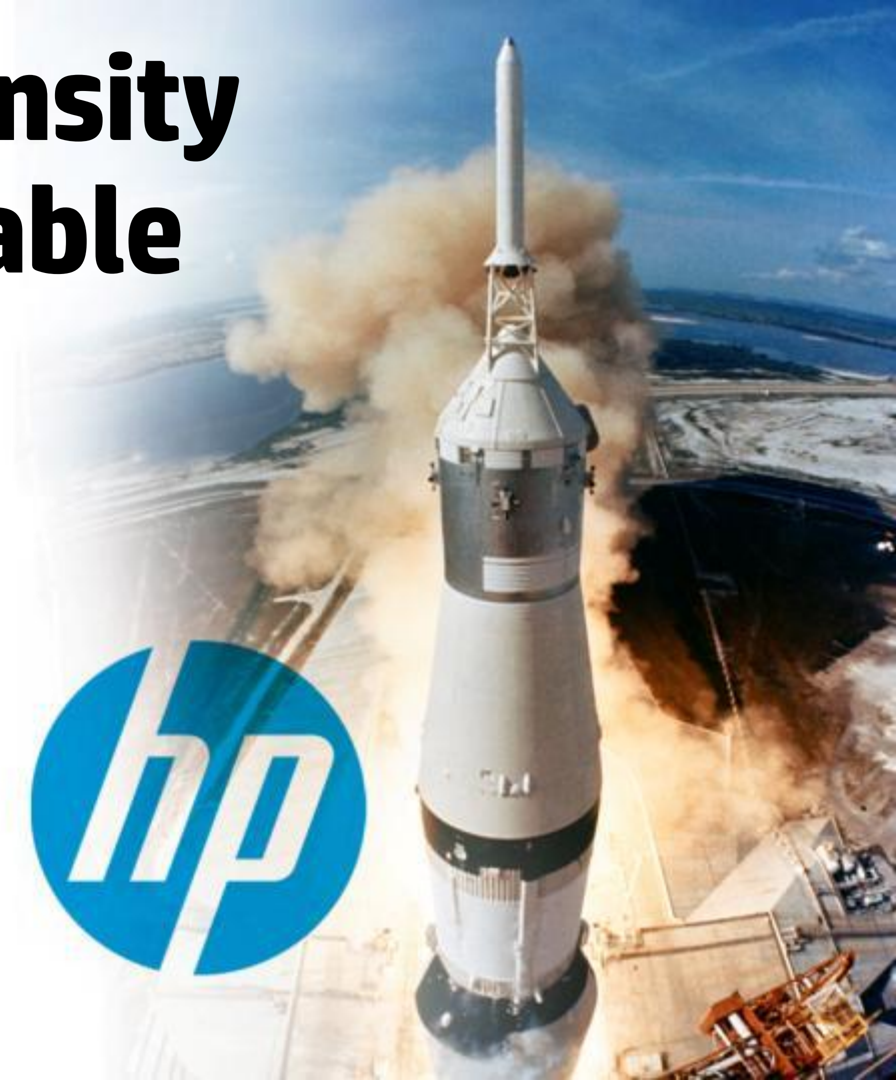
Low power, high density massive & manageable solutions

Project Moonshot

Sébastien Cabaniols

EMEA Presales & WW R&D consultant

Teratec, June 2013, FRANCE



IoT solutions drive new architecture requirements

Opportunity for competitive advantage serving more customers with unique offerings

Scale

Millions of apps and
billions of devices and users

Speed

Adapt at the speed of business
to gain competitive advantage

Specialized

Tailored & optimized for the
specific needs of each workload

The world's first software defined server

A modern architecture engineered for the new style of IT

Moonshot Architecture

10:1 Scaling *

Software Defined Servers

8x Efficiency *

Innovation Pace

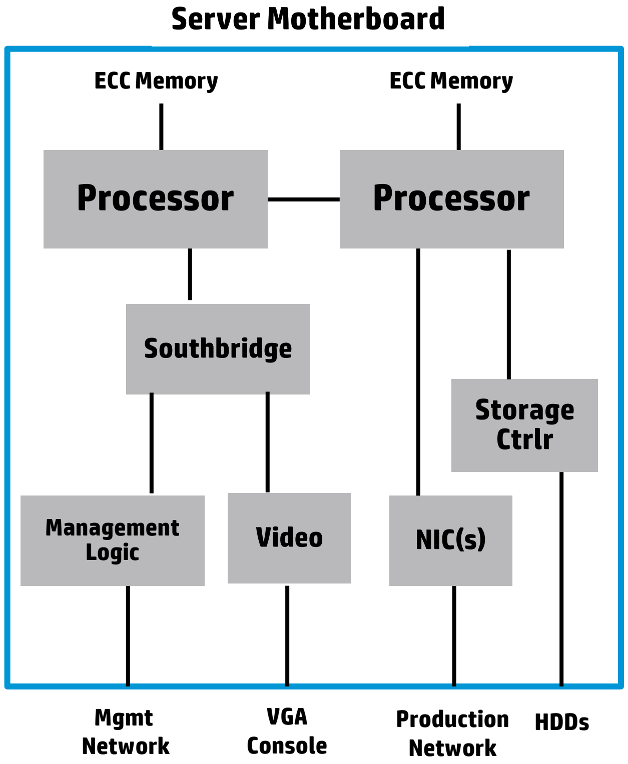
3x Faster *

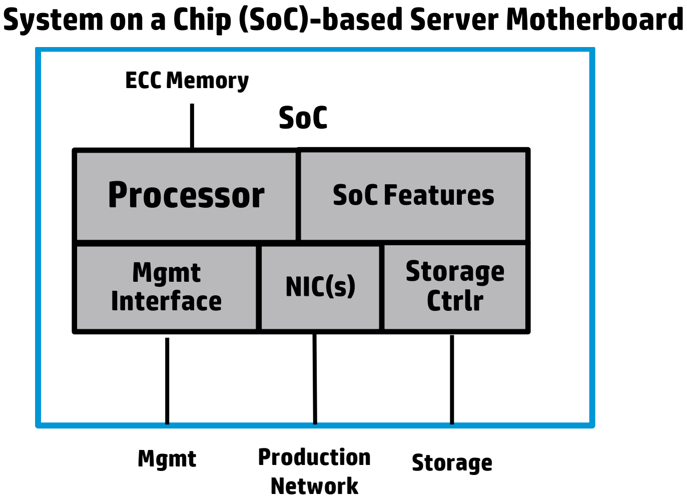
HP Moonshot System

*Source: HP internal research

HP Confidential until April 8, 2013

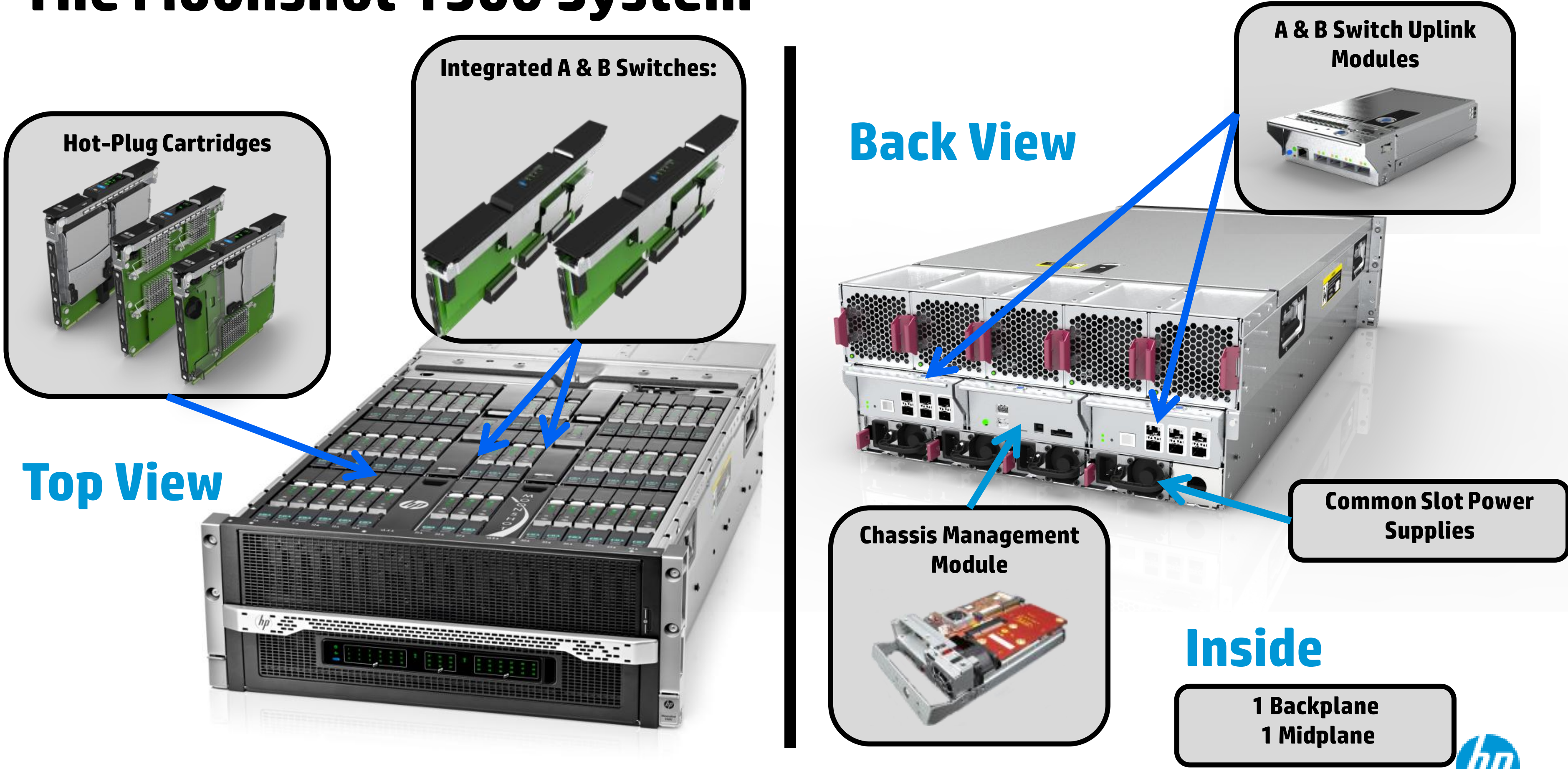
A New Era of Application-Focused Silicon





- Less general-purpose, more workload focused
- Dramatic reduction in power, cost, and space
- SoC vendors bring their own differentiated features and opportunities to disrupt markets

The Moonshot 1500 System



Hot-Plug Cartridges

Integrated A & B Switches:

Top View

Back View

A & B Switch Uplink Modules

Chassis Management Module

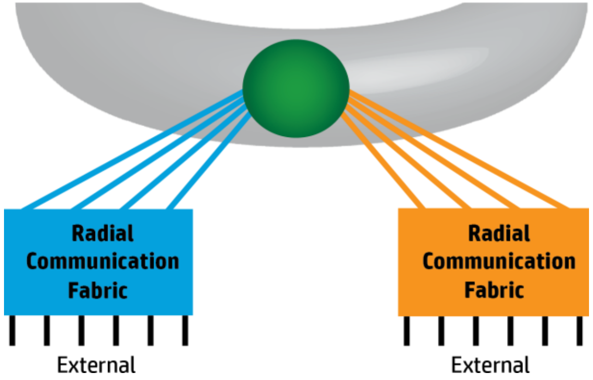
Common Slot Power Supplies

Inside

1 Backplane
1 Midplane

Moonshot Data Fabrics

Radial Communication



Radial Communication Fabric

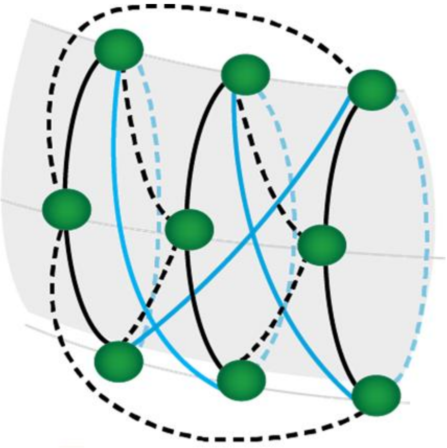
Radial Communication Fabric

External

External

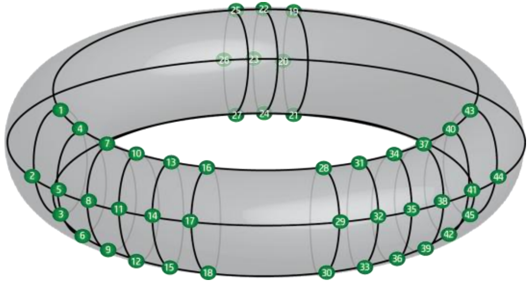
- High speed interfaces between each cartridge and two radial fabric slots; external connectivity

Proximal Array



- Five separate 3x3 proximal array fabrics within 2D Torus Mesh

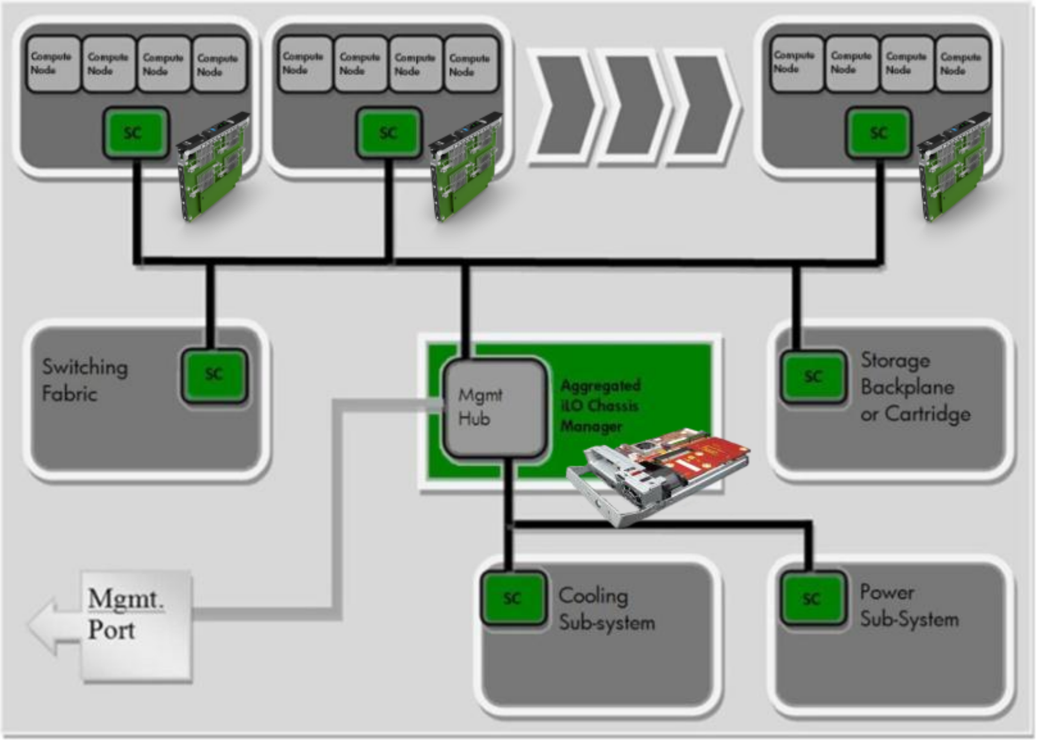
2D Torus Mesh



- High bandwidth cartridge-to-cartridge communication (North, South, East, West)

Management Fabric

Putting the hooks in place to allow for amazing flexibility!



- **Device neutral, low cost node solution**
- **Operates as 'brain' for chassis**
- **IPMI and Serial Console for each server**
- **Single Ethernet port gateway**
- **iLO Chassis Manager aggregates all to a common set of management interfaces**
- **SLAPM Rack Management spans rack or multiple racks**
- **True out of band firmware update services**

Insight CMU

HP Insight CMU = Cluster Management Utility

“CMU optimizes the TCO of compute farms”

- **CMU scaling specification: 4k nodes**
- **CMU has lots of industrial clusters in production with 2k/3k+ nodes**
  - **>100k compute nodes installed…**
  - **engineering, universities, government & research, energy…**
- **CMU has a strong presence in the TOP 500 (www.top500.org)**
- **CMU at customer site since 2000**
- **CMU has a strong & growing eco system with partner software (connectors)**

Insight CMU history

2013 - Moonshot support, ARM port in progress…

2011 - HP CMU joins the HP Insight family: HP Insight CMU

2010 - “Tsubame 2”, >1 PFlop cluster, 5th @ TOP500

2007 - Swedish gov, 6th @ TOP500

2004 - port to x86_64 Linux.

2002 - port to x86 & IA64 Linux / HPUX Itanium

2001 - port to Alpha Linux, 1600 servers commercial cluster

2000 - initial implementation for Tru64 Unix (Alphaserver)

Insight CMU project mindset

CMU provides the core functionalities for a compute farm

- runs any HP* server (even mix) / any Linux distribution (even mix)

- **independent** of many architectural aspects of the system:

  - interconnects / GPGPUs / CO-processors, IO-accelerators…
  - network topology (open cluster, guarded cluster, WAN…)
  - batch/job schedulers, MPI stacks, math libraries, compilers…

CMU is not a supercomputer software appliance

- most CMU systems delivered as “turn-key solutions”
- CMU can also be purchased standalone with support and manuals

Insight Cluster Management Utility Basics

- CMU is a single package running on the cluster head node (upgrade is trivial)

CMU mgt node can be an HA cluster (HP serviceguard, Redhat Cluster, SLES HA…)

install CMU mgt node in minutes (**see new cmu_mgt_config tool in 7.1**)

- provides an interactive CLI
- provides cmu_* commands as an API (for scripting)
- provides GUI client for single dashboard control
  - launch from a web page served from the head node (JAVA© webstart)
  - run on a local laptop/desktop
  - “user mode” for monitoring
  - “admin mode” for administration

HP Insight CMU ‘Three functional pillars’

Provisioning

- Simplified discovery
- Auto-Install { Kickstart, AutoYast… }
- Fast & scalable cloning engine
- Diskless support

Monitoring

- ‘At a glance’ view of entire system
- Customizable and HPC friendly
- 2D Instant View
- 3D Time View visualization
- History of performance metrics

Scalers

- GUI / CLI / API interfaces
- ‘One click’ access to servers
- cmudiff: Command broadcast & analysis

Proliant Rack Servers, Proliant Blade Servers, Proliant Moonshot Servers…

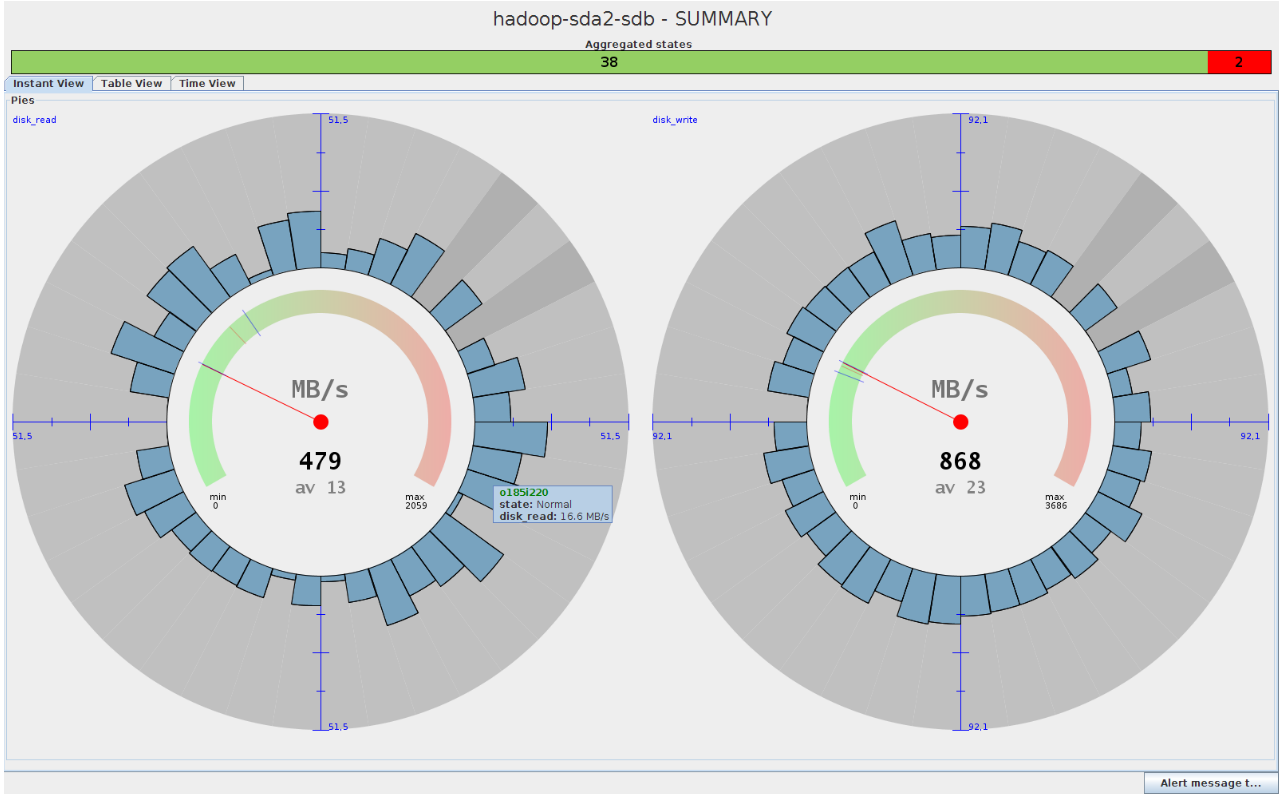
Insight CMU
3 unique features:

- InstantView,TimeView,Replay Engine
- cmudiff, command broadcast analyzer
- cmu API & connector program

Insight CMU: monitoring

InstantView
TimeView
Replay Engine

CMU Scalable ‘Instant View’



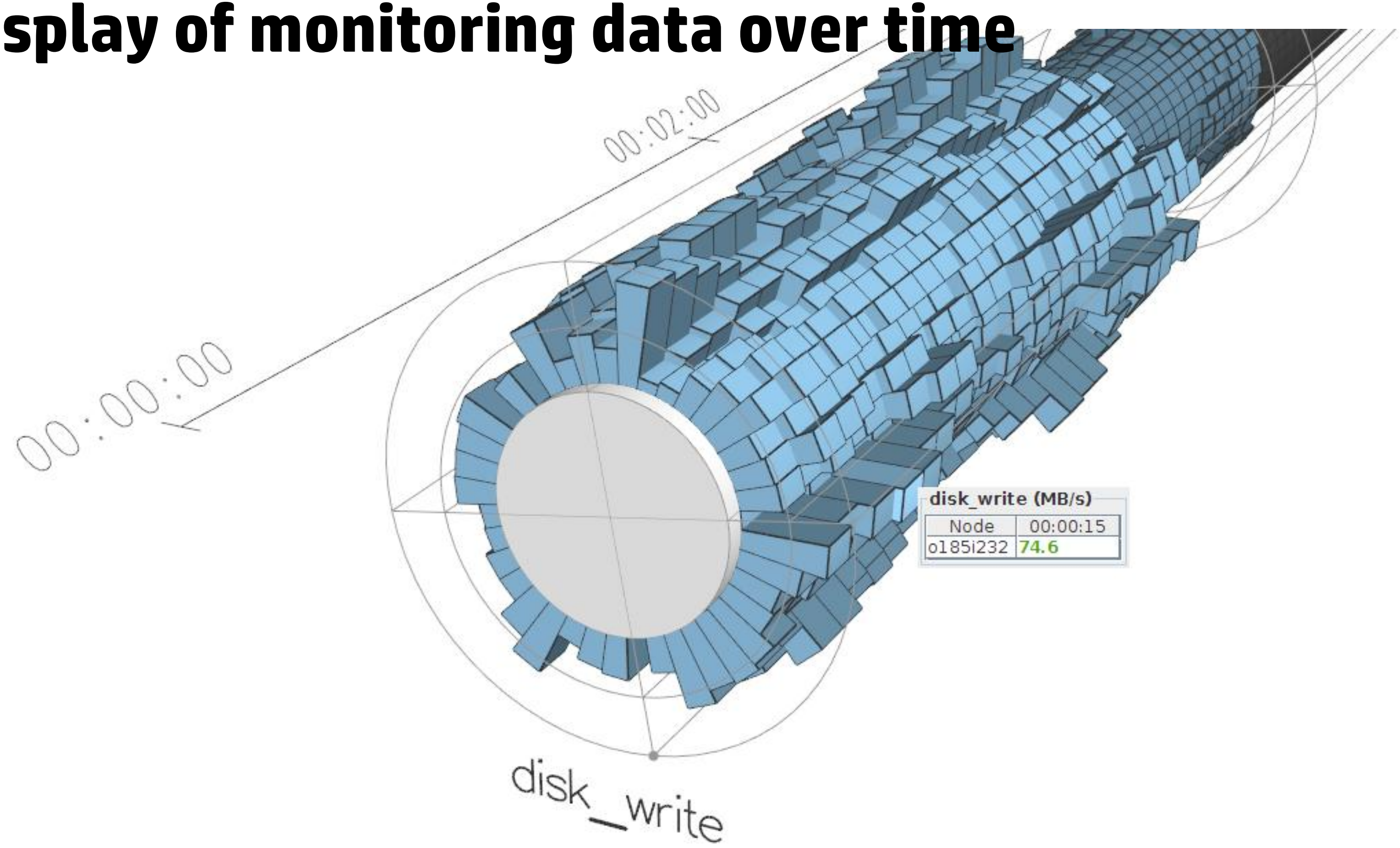
hadoop-sda2-sdb - SUMMARY

Aggregated states

38 | 2

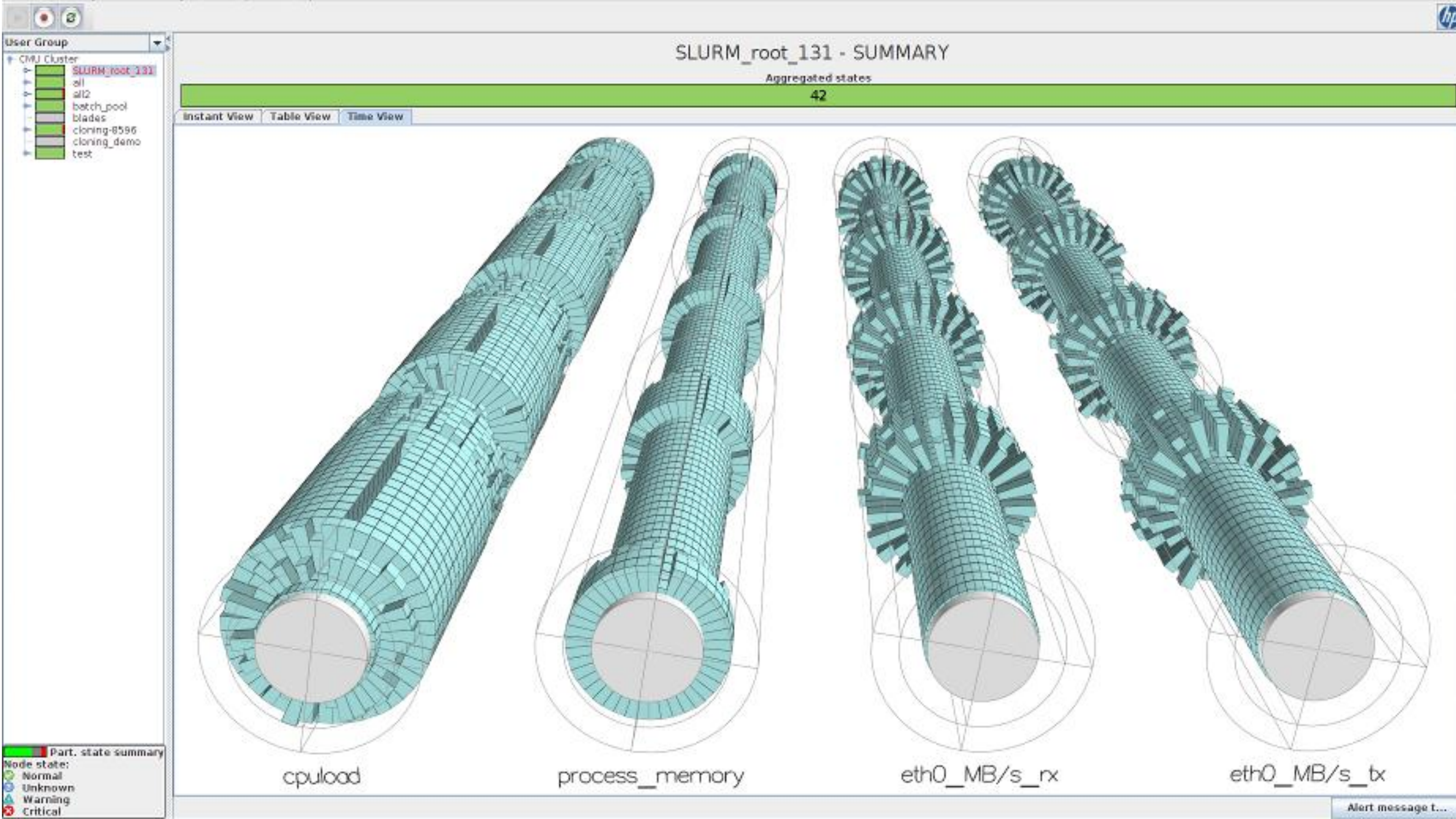
o185i220
state: Normal
disk_read: 16.6 MB/s

3D display of monitoring data over time



disk_write (MB/s)

Node	00:00:15
o185i232	74.6


User Group
CMU Cluster
SLURM_root_131
all
all2
batch_pool
blades
cloning-8596
cloning_demo
test
Part. state summary
Node state:
Normal
Unknown
Warning
Critical
SLURM_root_131 - SUMMARY
Aggregated states
42
Instant View
Table View
Time View
cpuload
process_memory
eth0_MB/s_rx
eth0_MB/s_tx
Alert message t...

CMU History Capability

- **store your CMU monitoring data, without compromises**

- at scale: 4096 nodes x 40 metrics x 5 secs/sample, for 3 years
  - no need of a dedicated, large/fast storage, fit in a few hundred gigs on a standard disk

- **access to your monitoring data efficiently**
  - within seconds even for very large datasets
  - for jobs up to ten thousands of server x hours (spec is 40000 server x hour now)
  - **retrieve with a user group interface allowing to retrieve a particular job by job-id.**

- **visualize with TimeView**
  - optimized client/server streaming only the necessary data

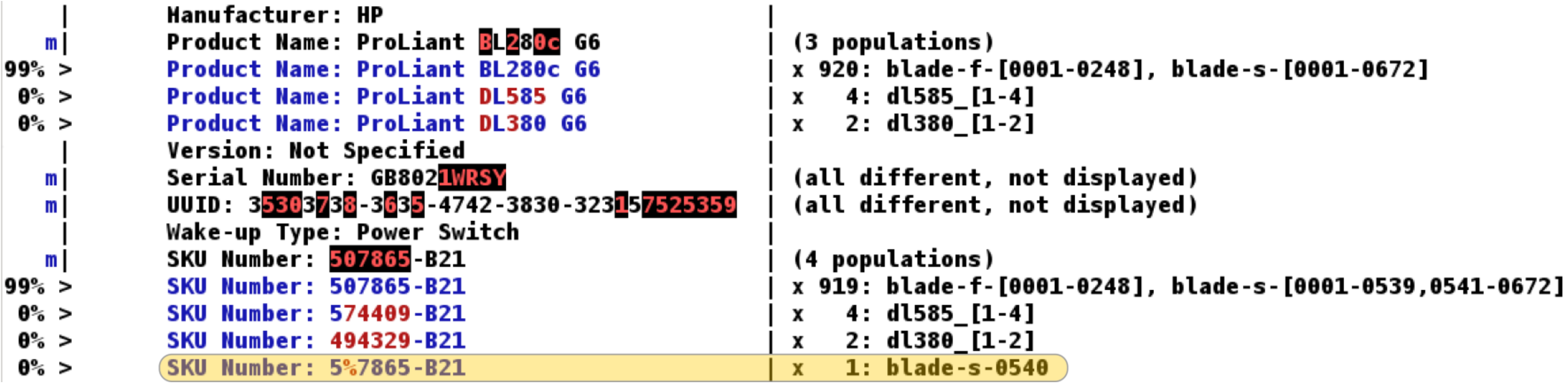
- **generate flat files (command line) to inject in another tool**

Insight CMU: cmudiff

« scaling the command line »

cmudiff

A real time data mining engine applied to cluster administration (pdsh post-processing)



```
    |     Manufacturer: HP                                | 
   m|     Product Name: ProLiant BL280c G6                | (3 populations)
99% >     Product Name: ProLiant BL280c G6                | x 920: blade-f-[0001-0248], blade-s-[0001-0672]
 0% >     Product Name: ProLiant DL585 G6                 | x   4: dl585_[1-4]
 0% >     Product Name: ProLiant DL380 G6                 | x   2: dl380_[1-2]
    |     Version: Not Specified                          |
   m|     Serial Number: GB8021WRSY                       | (all different, not displayed)
   m|     UUID: 35303738-3635-4742-3830-323157525359      | (all different, not displayed)
    |     Wake-up Type: Power Switch                      |
   m|     SKU Number: 507865-B21                          | (4 populations)
99% >     SKU Number: 507865-B21                          | x 919: blade-f-[0001-0248], blade-s-[0001-0539,0541-0672]
 0% >     SKU Number: 574409-B21                          | x   4: dl585_[1-4]
 0% >     SKU Number: 494329-B21                          | x   2: dl380_[1-2]
 0% >     SKU Number: 5%7865-B21                          | x   1: blade-s-0540
```

Example of 926 servers, running the ‘dmidecode’ command:

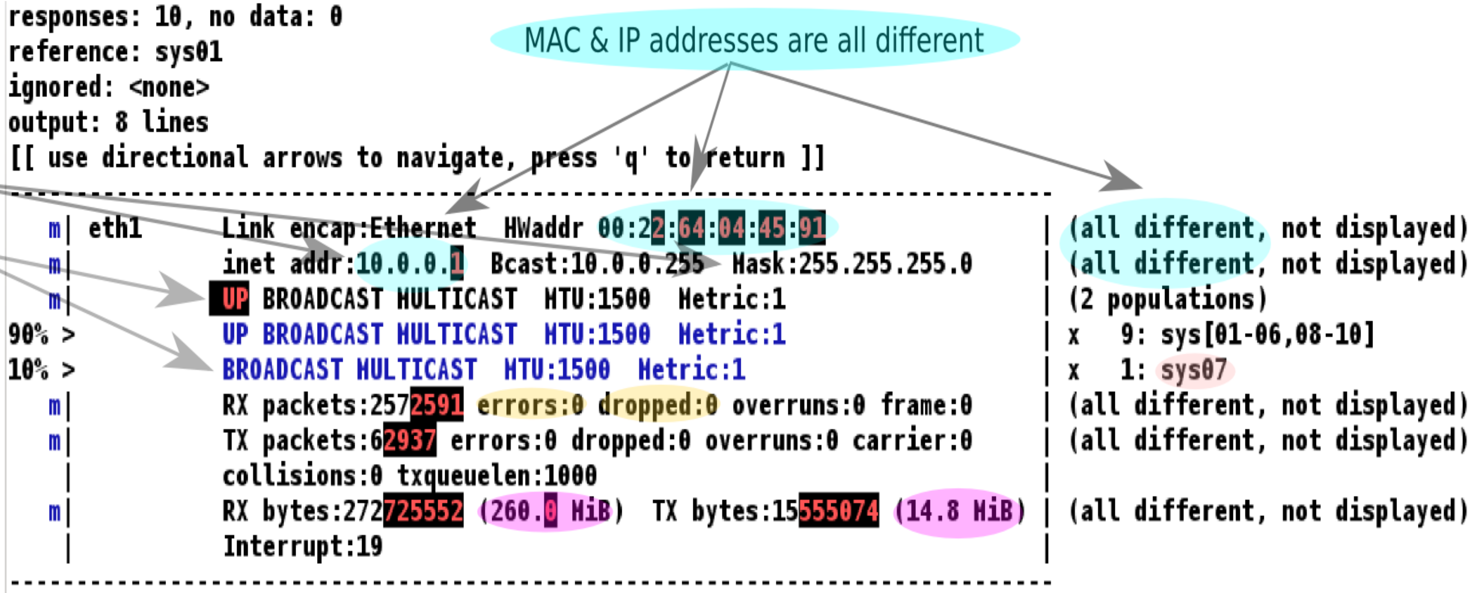
- 900k lines of text => (within seconds…) => 1918 lines report (~500x ratio)

=> all 5556 DIMMS in the cluster are identical in speed/size/slotting/model

=> __one__ unexpected ROM flash anomaly (highlighted in yellow)

cmudiff

Exemple: running the ‘ifconfig’ command



```
responses: 10, no data: 0
reference: sys01
ignored: <none>
output: 8 lines
[[ use directional arrows to navigate, press 'q' to return ]]
----------------------------------------------------------------------
   m| eth1      Link encap:Ethernet  HWaddr 00:22:64:04:45:91              | (all different, not displayed)
   m|           inet addr:10.0.0.1  Bcast:10.0.0.255  Mask:255.255.255.0  | (all different, not displayed)
   m|           UP BROADCAST MULTICAST  MTU:1500  Metric:1                 | (2 populations)
90% >           UP BROADCAST MULTICAST  MTU:1500  Metric:1                 | x   9: sys[01-06,08-10]
10% >           BROADCAST MULTICAST  MTU:1500  Metric:1                    | x   1: sys07
   m|           RX packets:2572591 errors:0 dropped:0 overruns:0 frame:0   | (all different, not displayed)
   m|           TX packets:62937 errors:0 dropped:0 overruns:0 carrier:0   | (all different, not displayed)
    |           collisions:0 txqueuelen:1000                               |
   m|           RX bytes:272725552 (260.0 MiB)  TX bytes:15555074 (14.8 MiB) | (all different, not displayed)
    |           Interrupt:19                                               |
----------------------------------------------------------------------
```

MAC & IP addresses are all different

all interfaces are configured in 10.0.0.0/255.255.255.0

"sys07" is the only system not reporting eth1 as "UP"

all systems reported 0 errors, 0 drops...

systems transferred similar volumes of data

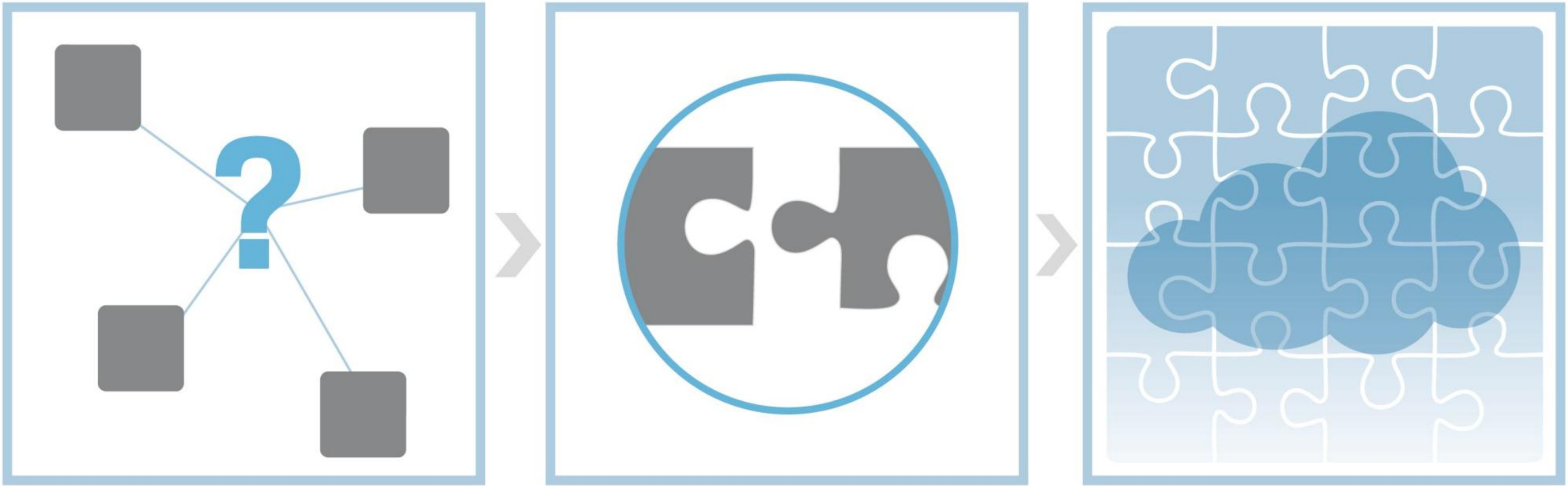
i.e Received ~ 260 Mib

and Transmittes ~ 14.8 Mib

Insight CMU:

API & connectors

CMU Connectors



From a variety of great tools…

- different vendors,
- different development cycles,
- different interfaces…

… to a fully integrated solution

- Joint development HP & Partners
- HP & Partners validated, tested
- distributed, maintained by partners

Insight CMU Connectors

- CMU UFM Connector (Mellanox)
- CMU PBS PRO Connector (Altair)
- CMU Moab Connector (Adaptive)

- HP Cloudera Hadoop appliance: CMU Ganglia Connector…
- HP Matrix CMU CloudMap

Thank You