

## CEP4 Update

October 26, 2016



A server room with blue and green lighting. In the foreground, several Mellanox network cables are visible, some with labels like 'Mellanox' and '2015-06'. The background shows server racks with glowing green lights.

# The road we've traveled: Timeline

# CEP4

## Timeline

- **August 2015:** Target system delivery
- **October 2015:** Realized hardware delivery; Start acceptance
- **January 2016:** Agreement on performance results: 4 extra storage servers needed; Start system configuration by CIT
- **April 2016:** Additional storage servers delivered; CEP4 support in specification & control systems
- **May/June 2016:** CEP4 support in pipelines, Cobalt, new scheduler; Internal CEP4 training workshops; CIT role changed from installation to support; Start commissioning
- **July 2016:** CEP4 clean-up tool delivered; Problematic data server repaired
- **August 2016:** Last activity DELL (configuration)



A server room with blue and green lighting. In the foreground, several Mellanox network cables are visible, some with labels like 'Mellanox' and '2015-06'. The background shows server racks with glowing green lights.

# The bare facts: System description

- 2 head nodes
  - 3.6 TB local storage
  - Failover configuration
- 3.1 PB global storage
  - Available on all nodes + lexar003/4 (LTA ingest)

System	Compute	Memory
head01.cep4 head02.cep4	12 cores @ 1.6 GHz	128 GB
50x cpuXX.cep4	24 cores @ 2.5 GHz	256 GB
4x gpuXX.cep4	16 cores @ 2.4 GHz 4x Tesla K40C	320 GB

- Observational throughput Cobalt -> CEP4:

Mode	CEP2	CEP4
BF superterp	<37 Gbps	>110 Gbps
Correlator	~40 Gbps	~69 Gbps

- Pipeline throughput TBD. CEP2 -> CEP4:
  - Batch scheduling -> better utilization
  - 4.8x more FLOPs
  - 0.8x memory bandwidth
  - 1.5x disk speed (but higher latency!)





# A bumpy ride: Challenges encountered

- CEP4 is not a drop-in replacement for CEP2
  - New technologies introduced
  - Development & testing require representative environment
- CEP4 is not a standard HPC cluster
  - Combination of real-time & offline processes
  - Part of an operational instrument



- IO performance acceptance
  - Vendor initial design based on reference implementation
  - Next time: representative LOFAR benchmark?
- Lustre behavior different from local disks
  - Higher throughput but also higher latency: random access significantly slower (e.g. generation inspection plots)
  - Large scale filesystem operations expensive: introduced "Robinhood" file system monitoring tool but synchronization challenging although it is an accepted tool. LOFAR file generation mechanisms atypical?
- Technology maturity
  - Bug in Slurm/Docker/Kernel cgroup handling

- Cobalt data loss
  - Data loss observed during start-up of imaging observations and throughout high throughput observation (LOTAAS)
  - Related to simultaneous offline process activity
  - First update OS, docker & configuration (CIT)



A server room with blue and green lighting. In the foreground, several Mellanox network cables are visible, some with labels like 'Mellanox' and '2015-06'. The background shows server racks with glowing green lights.

**A bright future:  
New features**

# CEP4

## Brings us many new things



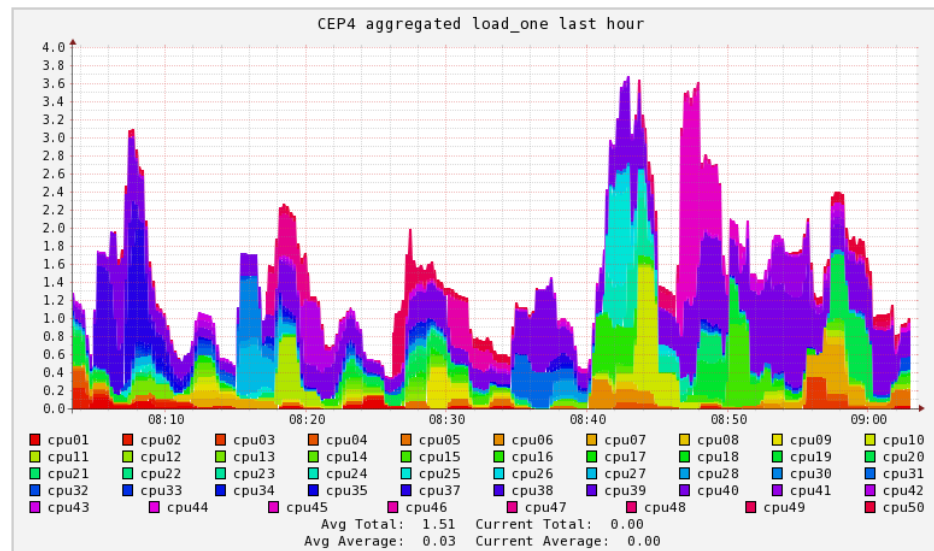
- Lustre: global filesystem
  - Easier data management & improved reliability/flexibility
- SLURM: batch scheduling
  - Saves significant operator effort (automation & maintenance)
  - Collects job statistics: characterization & optimization
- Docker: software configuration management
  - Improves version handling/availability (continuity, rollback)
  - Easier deployment of user-provided tools
- C++11: compatibility with latest Linux & compilers
- SupervisorD: process management
- Ganglia: cluster monitoring



# Monitoring CEP4 Utilization (Ganglia)

- Many metrics: CPU/network/memory/SLURM/...
- Extensive querying: per day/week/month/..., graph/cvs/json

Stacked Graph - load\_one



CEP4 load\_one last hour sorted by name

Metric:  Show Hosts Scaled:    Size:  Columns:  (0 = metric + reports)

Show only nodes matching:  Filter Max graphs to show:  Sorted:



...and already set in motion before,  
but realized on CEP4

- Services Framework: more flexible interaction between components
  - Modular: easier maintenance & development
  - Open interfaces: easier access to information



# ...and already set in motion before, but realized on CEP4



- ResourceAssigner: a new Scheduler
  - Real-time scheduling: ready for responsive telescope
  - Web-based
  - Faster & interactive

The screenshot displays the ResourceAssigner interface, which is used for scheduling observations. It features a table of observation details on the left and a Gantt chart on the right.

**Table Headers:** Name, Project, Start, End, Duration, Status, Info, Type, Size, Group ID, Mem ID, SAS ID, RADR ID, Chatter...

**Table Content (Sample Rows):**

Name	Project	Start	End	Duration	Status	Type	Size	Group ID	Mem ID	SAS ID	RADR ID	Chatter
B0820-02	LC6_028	2016-10-26 06:00:00	2016-10-26 06:15:00	00:15:00	finished	observation		716043	716044	547979	24430	CEP2
CV/PSR1022...	IPS_Commissioni...	2016-10-26 08:45:00	2016-10-26 08:55:00	00:10:00	scheduled	observation		729196	729197	555099	31311	CEP2
J1400-1438	LC6_028	2016-10-26 11:28:00	2016-10-26 11:28:00	00:00:00	scheduled	observation		716043	716047	547983	24431	CEP2
LOTAAS-P07...	LTS_004	2016-10-26 18:11:00	2016-10-26 19:11:00	01:00:00	approved	observation		718806	718807	549961	25725	CEP2

**Gantt Chart:** The Gantt chart shows the timeline of observations from October 26, 2016, to October 27, 2016. It includes bars for various observation types, such as 'ALL OBSERVATIONS (65.8%)', 'ALL PIPELINES (34.4%)', 'LTS\_009 observation', 'IPS\_Commissioning pipeline', 'CVIS', 'LTAAS-P090/P10LP', and 'LOTA' observations. The chart is color-coded by project or observation type.



A server room with blue lighting and Mellanox network cables. The image shows a dense array of server racks with numerous blue and black network cables plugged into the front panels. The cables are labeled with 'Mellanox' and other technical specifications. The background is filled with server racks and glowing green lights, creating a high-tech, data-center atmosphere.

**Fin  
Questions?**