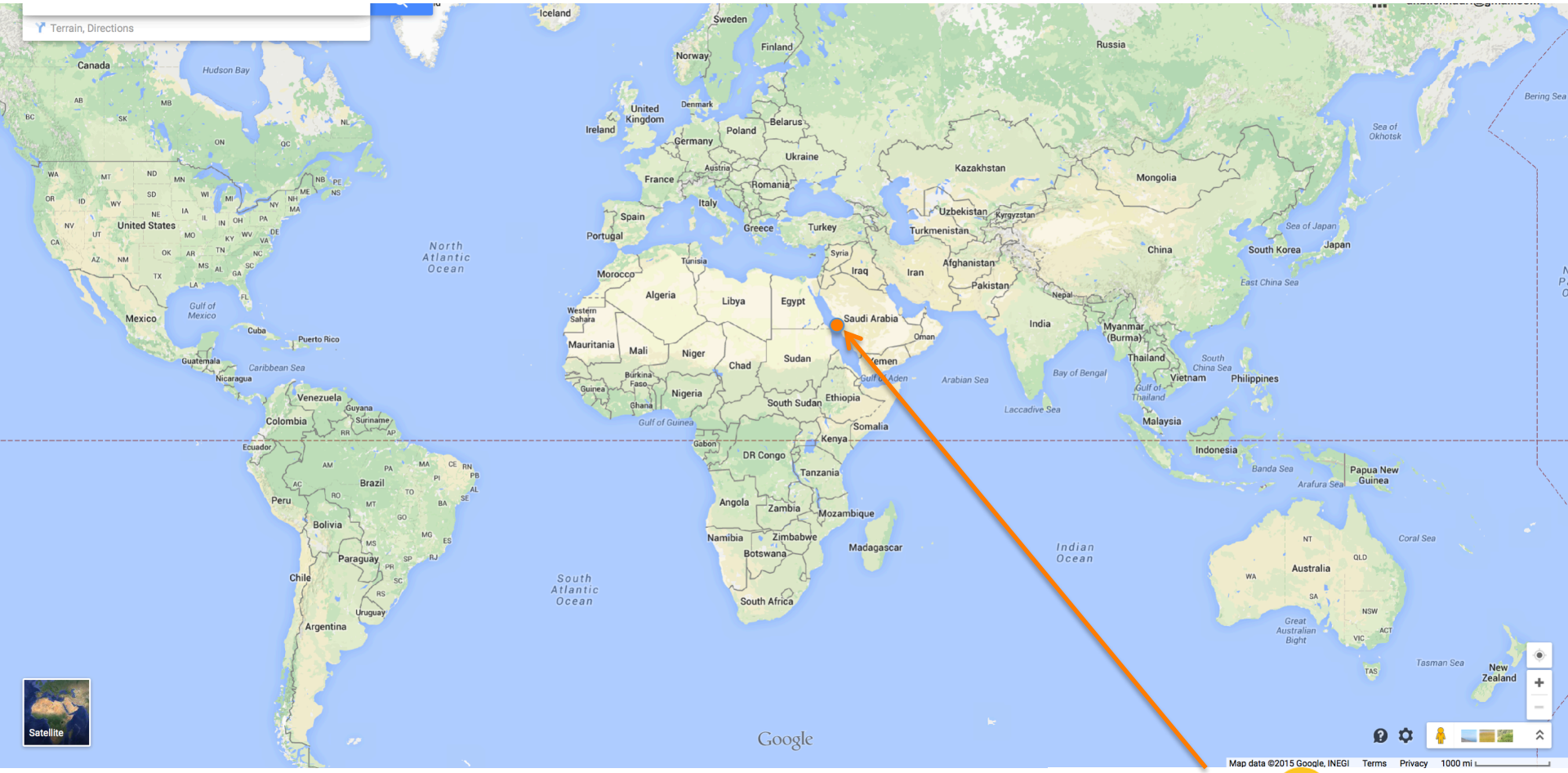
Overview of the KAUST's Cray X40 System – Shaheen II

Bilel Hadri, Samuel Kortas, Saber Feki, Rooh Khurram, Greg Newby

KAUST Supercomputing Laboratory

Where is KAUST ?



KAUST is located on the shores of the Red Sea, in Saudi Arabia

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

KAUST

King Abdullah University of Science and Technology (KAUST) is an international, graduate research university in Saudi Arabia, dedicated to advancing science and technology through interdisciplinary research, education and innovation. KAUST has 3 main divisions:

- Biological Science and Engineering (BESE)
  - Bioscience
  - Environmental Science & Engineering
  - Marine Science
  - Plant Science
- Computer, Electrical & Mathematical Science & Engineering (CEMSE)
  - Applied Mathematics and Computational Science
  - Computer Science
  - Electrical Engineering
- Physical Science & Engineering (PSE)
  - Chemical and Biology Engineering
  - Chemical Science
  - Earth Science and Engineering
  - Material Science and Engineering
  - Mechanical Engineering

KAUST Research Centers

- The University has 11 Research Centers
  1. Advanced Membranes and Porous Materials Center (AMPMC)
  2. Catalysis (KCC)
  3. Clean Combustion (CCRC)
  4. Computational Bioscience (CBRC)
  5. Center for Desert Agriculture (CDA)
  6. Extreme Computing Research Center (ECRC)
  7. Red Sea Research Center (RSRC)
  8. Solar and Photovoltaics Engineering Research Center (SPERC)
  9. Upstream Petroleum Engineering Center (UPEC)
  10. Visual Computing Center (VCC)
  11. Water Desalination and Reuse (WDRC)

- The Academic Divisions and Research Centers support the University's research mission by bringing together faculty members, researchers, and graduate students from across the disciplines. Together, they leverage the interconnectedness of science and engineering and develop interdisciplinary approaches to fundamental and goal-oriented research.

Core Labs

- The Core Labs and Major Facilities offer state-of-the art research equipment operated by more than one hundred expert staff scientists to support KAUST's research community.

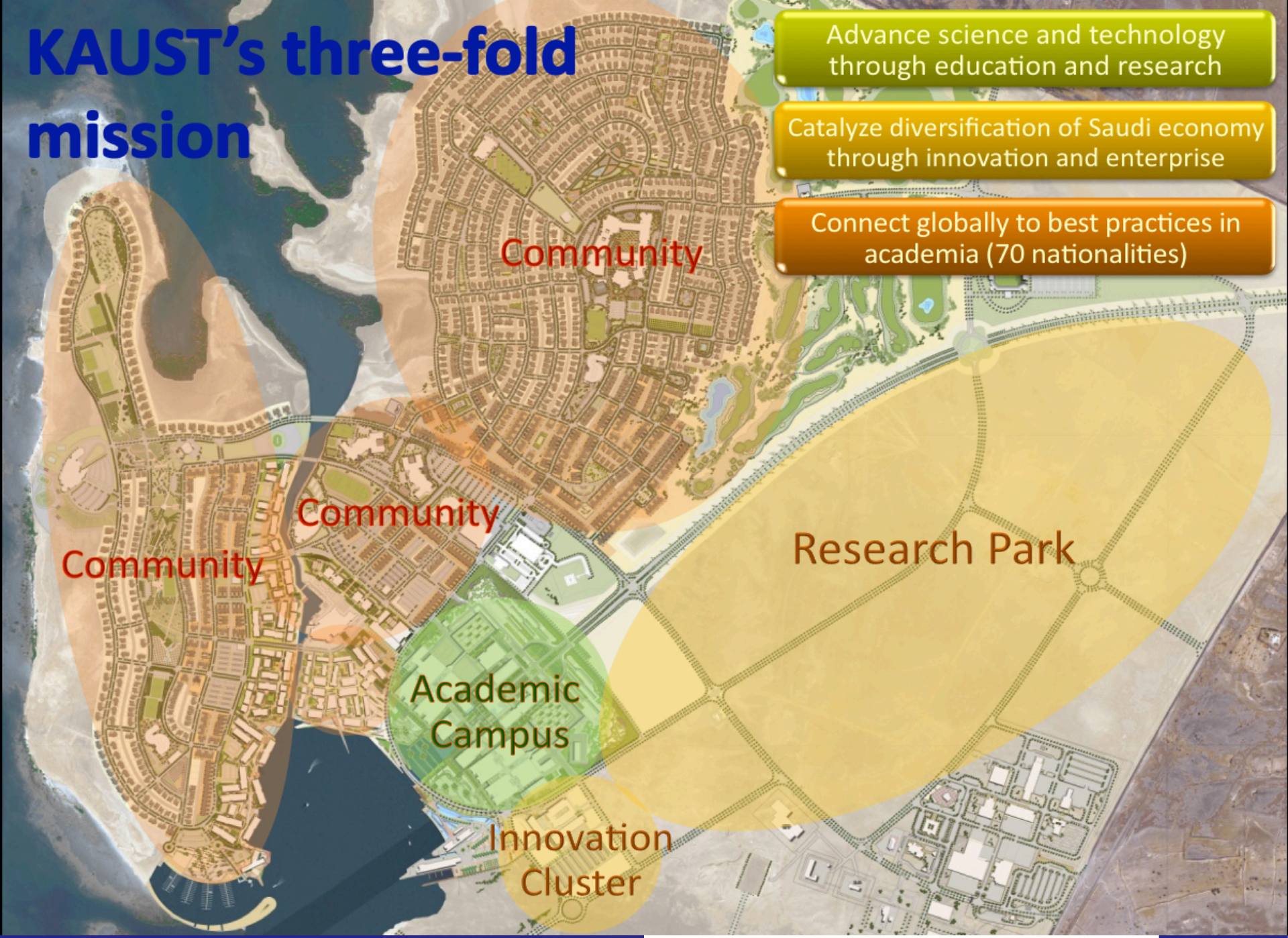
- The University has 8 Core Labs and Major Facilities:
  ① Advanced Nanofabrication and Thin Film Core Lab
  ② Analytical Core Lab
  ③ Biosciences Core Lab
  ④ Costal and Marine Resources Core Lab
  ⑤ Imaging and Characterization Core Lab
  ⑥ **Supercomputing Core Lab (KSL)**
  ⑦ Visualization Core Lab
  ⑧ Central Workshop

KAUST's three-fold mission

KAUST's three-fold mission

KAUST's three-fold mission

KAUST's three-fold mission



جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

KAUST city

KAUST is like a city (36 km2 - 14 sq mi):

- Campus
- K12 schools
- Daycare centers
- Restaurants, cafes, fine dining
- Cinemas, concert theatre
- Bank, post office, travel agent
- Beauty salon, dry cleaner
- Golf
- Supermarket
- Student Housing
- Faculty Villas and Staff Housing
- Recreational Facilities
- Stadium
- Beaches
- Hotels
- Security and Fire Protection
- Health Clinic and Heliport

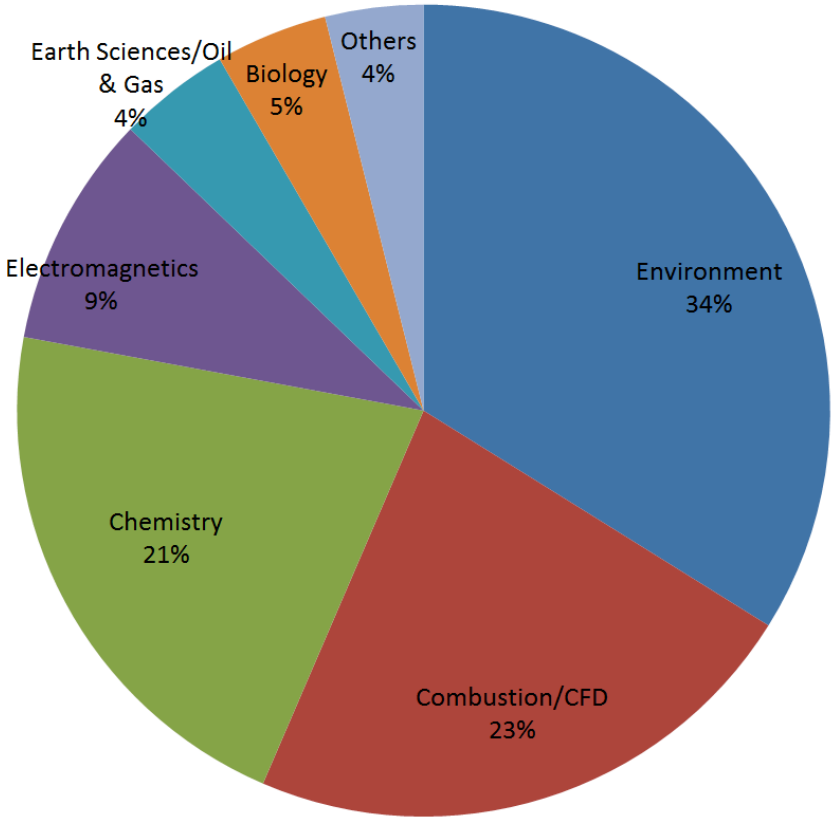
More info on http://www.kaust.edu.sa/

KAUST Applications

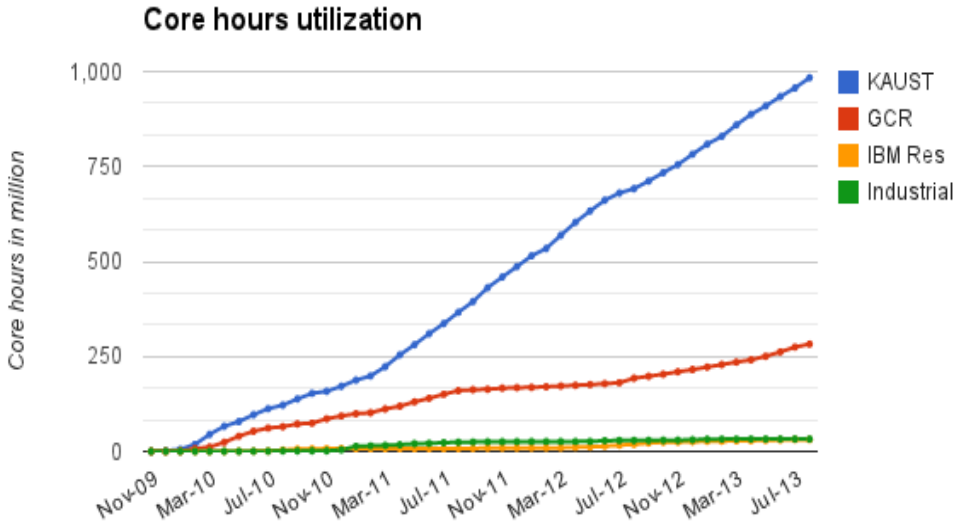
Science Area	Codes
Atmospheric Modeling	WRF,WRF-Chem,HIRAM
Ocean Modeling	WRF, MITgcm
Combustion	NGA, S3D
CFD/Plasma	Plasmoid – in house code
Biology	In-house genomic motif identification code
Earthquake Seismology	SORD , SeisSol, SPECFEM_3D_GLOBE
Electromagnetism	In house explicit code
Big Data/	Mizan - in house code (Analysis of Large Graphs)
Chemistry	VASP, LAMMPS, Gaussian, WEIN2k,Quantum Espresso
Seismic imaging/Oil & gas	In house 3D reverse time migration code

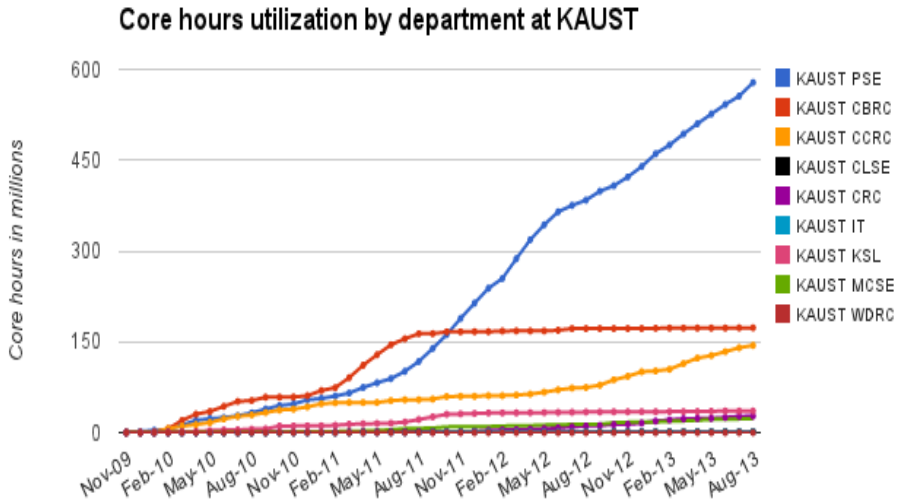
More details http://www.hpc.kaust.edu.sa/sc14/presentations/

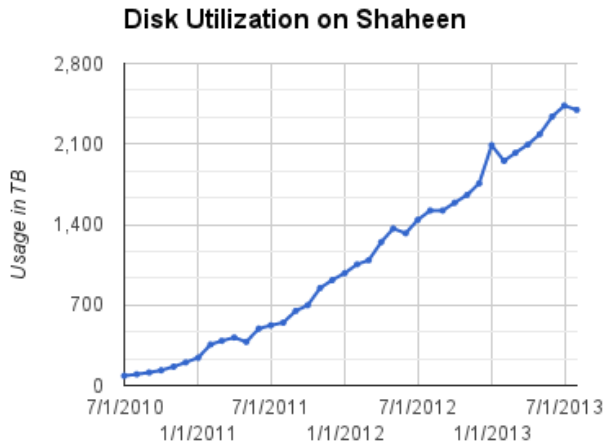
Shaheen I Utilization by Science Area (2009-2015)



Shaheen I Utilization

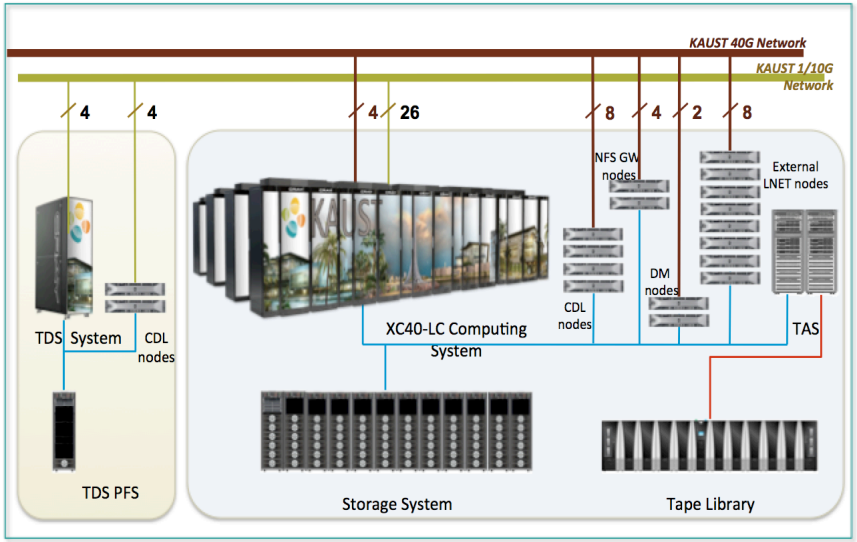






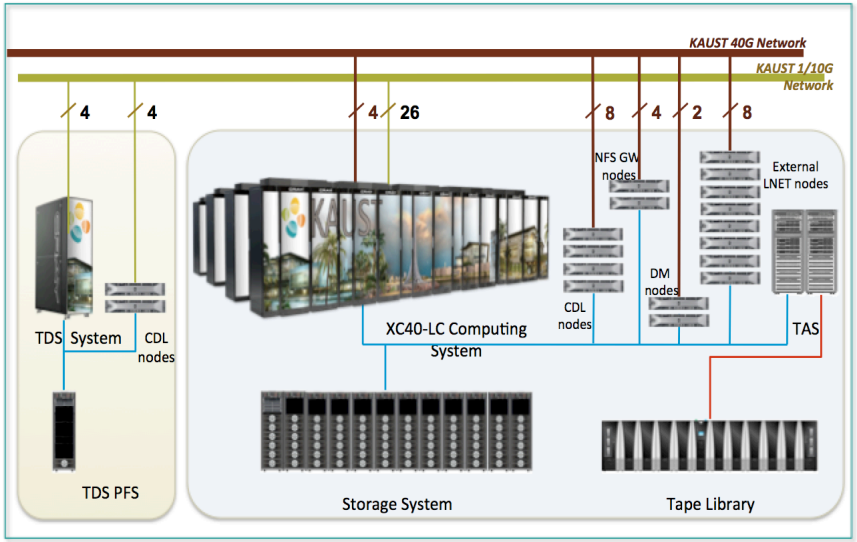
Shaheen II Overview

COMPUTE	Node	Processor type: Intel Haswell	2 CPU sockets per node, 16 processors cores per CPU, 2 .3GHz
		6174 Nodes	197,568 cores
		128 GB of memory per node	Over 790 TB total memory
	Power	Up to 2.8MW	Water Cooled
	Weight/Size	More than 100 metrics tons	36 XC40 Compute cabinets, plus disk, blowers, management , etc..
	Speed	7.2 Pflop/s speak theoretical performance	Over 5 Pflop/s sustained LINPACK
	Network	Cray Aries interconnect with Dragonfly topology	57% of the maximum global bandwidth between the 18 groups of two cabinets.



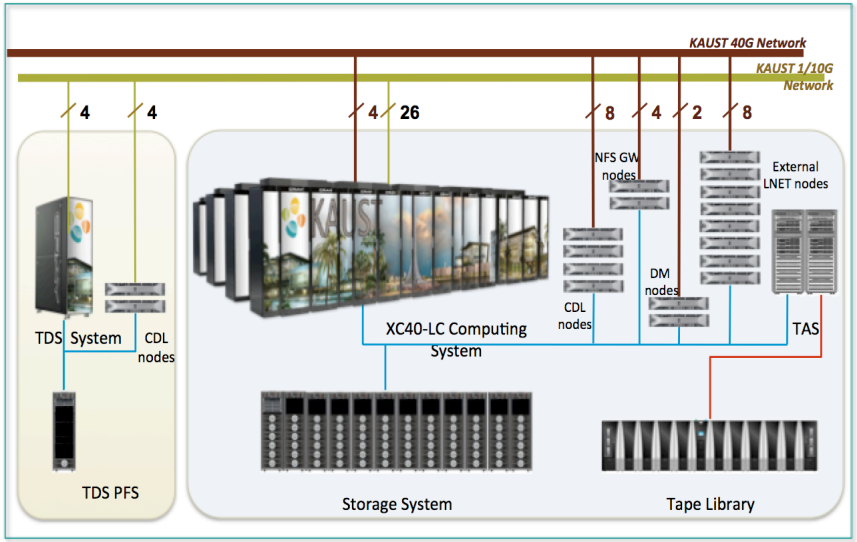
Shaheen II Overview

COMPUTE	Node	Processor type: Intel Haswell	2 CPU sockets per node, 16 processors cores per CPU, 2 .3GHz
		6174 Nodes	197,568 cores
		128 GB of memory per node	Over 790 TB total memory
	Power	Up to 2.8MW	Water Cooled
	Weight/Size	More than 100 metrics tons	36 XC40 Compute cabinets, plus disk, blowers, management , etc..
	Speed	7.2 Pflop/s speak theoretical performance	Over 5 Pflop/s sustained LINPACK
	Network	Cray Aries interconnect with Dragonfly topology	57% of the maximum global bandwidth between the 18 groups of two cabinets.
STORE	Storage	Sonexion 2000 Lustre appliance	17.6 petabytes of usable storage. Over 500 GB/s bandwidth
	Burst Buffer	DataWarp	Solid Sate Devices (SDD) fast data cache. Over 1 TB/s bandwidth, (delivery September 2015)
	Archiving	Tiered Adaptive Storage (TAS)	Hierarchical storage with 200 TB disk cache and 20 PB of tape storage, using a spectra logic tape library. (can expand up to 100 PB)



Shaheen II Overview

COMPUTE	Node	Processor type: Intel Haswell	2 CPU sockets per node, 16 processors cores per CPU, 2.3GHz
		6174 Nodes	197,568 cores
		128 GB of memory per node	Over 790 TB total memory
	Power	Up to 2.8MW	Water Cooled
	Weight/Size	More than 100 metrics tons	36 XC40 Compute cabinets, plus disk, blowers, management , etc..
	Speed	7.2 Pflop/s speak theoretical performance	Over 5 Pflop/s sustained LINPACK
	Network	Cray Aries interconnect with Dragonfly topology	57% of the maximum global bandwidth between the 18 groups of two cabinets.
STORE	Storage	Sonexion 2000 Lustre appliance	17.6 petabytes of usable storage. Over 500 GB/s bandwidth
	Burst Buffer	DataWarp	Solid Sate Devices (SDD) fast data cache. Over 1 TB/s bandwidth, (delivery September 2015)
	Archiving	Tiered Adaptive Storage (TAS)	Hierarchical storage with 200 TB disk cache and 20 PB of tape storage, using a spectra logic tape library. (can expand up to 100 PB)
ANALYZE	Analyzing	Urika - GD	2TB of global shared-memory, 64 Threadstorm4 processors with 128 hardware threads per processor Over 75 TB of Lustre PFS



Shaheen II: Software ecosystem

- Shaheen II system is tightly integrated with compute, storage and data analytics solutions
  - Increases the difficulty to manage efficiently the hardware and software.
- KSL supports hundreds of third party packages.
  - Need to keep the installations consistent, up-to-date and providing reproducible performance and correctness of the results.
- Solution : Shaheen II software ecosystem:
  - monitoring,
  - software management,
  - regression tools
  - Efficient scheduling manager
- Strategy: not to re-invent the wheel.

Monitoring

- Monitor SW usage :
  - What are the most linked, compiled and executed applications your HPC system? How do we find who is using deprecated software or versions with bugs?....
  - Solution: XALT:
    - improves the functionalities of ALTD by tracking users, codes and environments. It collects job-level and link-time level data and subsequent analytics automatically and transparently.
    - Already in place on many centers, more than dozen in US and Europe (NICS, ORNL, CSCS, NERSC,NSCA,, TACC, KAUST …)
    - Great Tutorial 2B by M. Fahey and R. Budiardja (easy instructions to install !)
    - At KAUST, ported ALTD on BG/P in 2013
      - Helped to extract most used libraries and applications with **real metrics**
      - Assisted in the design of benchmarks used for procurements benchmarking
      - Detected some bloopers (ref. blas/lapack, mpirun –np 4 ./config –prefix=…!!!)

Monitoring

- Monitor SW usage :
  - What are the most linked, compiled and executed applications your HPC system? How do we find who is using deprecated software or versions with bugs?....
  - Solution: XALT:
    - improves the functionalities of ALTD by tracking users, codes and environments. It collects job-level and link-time level data and subsequent analytics automatically and transparently.
    - Already in place on many centers, more than dozen in US and Europe (NICS, ORNL, CSCS, NERSC,NSCA,, TACC, KAUST …)
    - Great Tutorial 2B by M. Fahey and R. Budiardja (easy instructions to install !)
    - At KAUST, ported ALTD on BG/P in 2013
      - Helped to extract most used libraries and applications with **real metrics**
      - Assisted in the design of benchmarks used for procurements benchmarking
      - Detected some bloopers (ref. blas/lapack, mpirun –np 4 ./config –prefix=…!!!)
- Monitor I/O performance
  - Checking the performance of the PFS.
  - Solution: Darshan
- Monitor Power Usage
  - SLURM Native with on-the-fly dynamic steering of the frequencies of every running jobs.

SWTools

- To maintain infrastructure for software management of the third-party installation, SWTools has been put in place.

/base	/machine	/appli	/version	/build
/sw	/tds	/gsl	/1.15	cnl5.2_cce8.3.2
				cnl5.2_intel15.2.2
				cnl5.2_gnu4.9.1

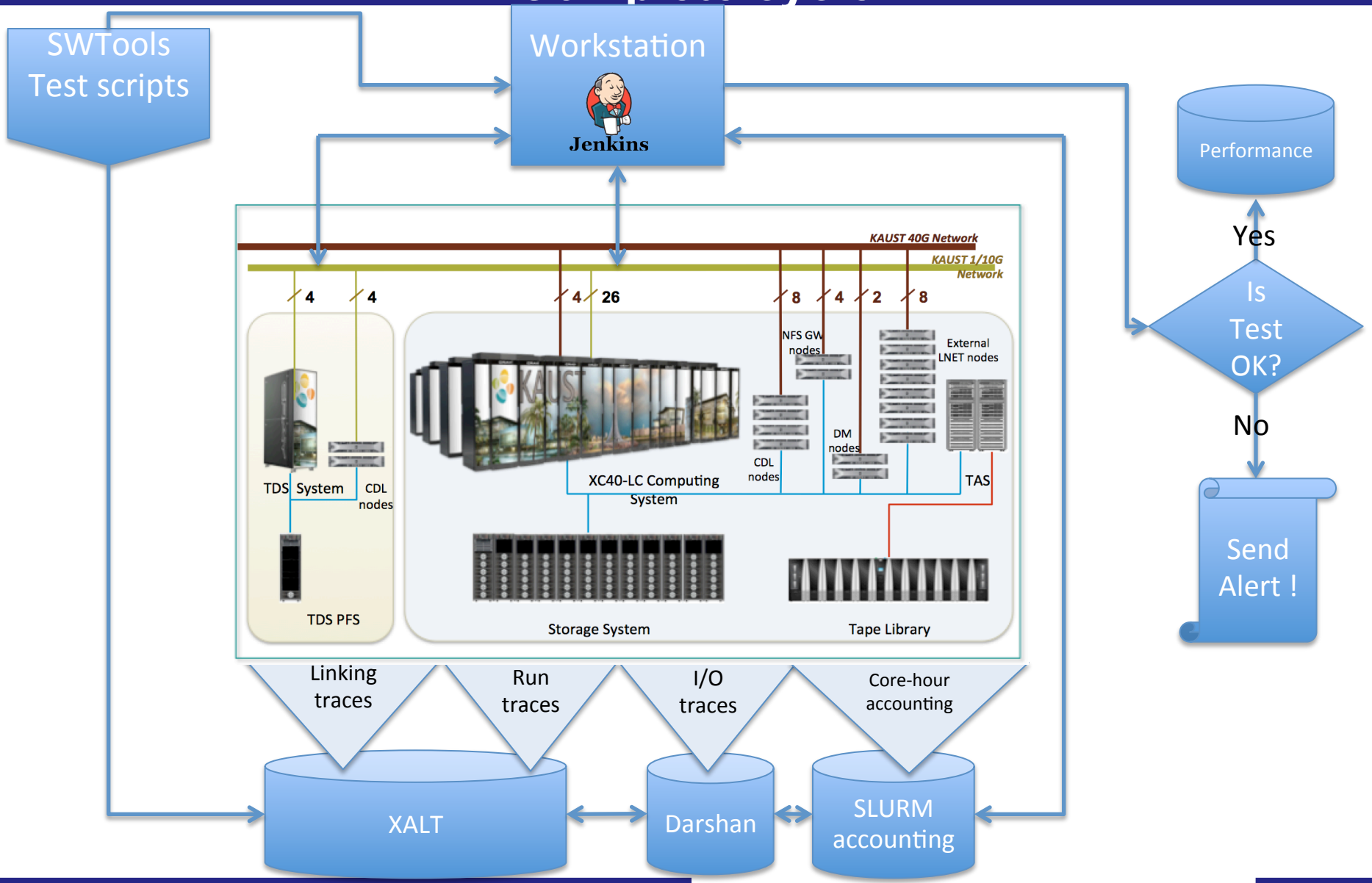
- Standardize workflow for software installations in order to get:
  - a clear documentation on installations
  - an automated building, linking and testing of installations
  - an inventory of currently installed software
  - an easily maintainable installation
  - an automate generation of many user documents

- Each Software installation needs to have test case (small and big) used
  - Validating installation
  - User documentation
  - For regression

Jenkins

- Along with the software management, KSL developed an early version of an automatic testing tool for the non-regression testing
- Needed to detect errors and/or performance degradation, allowing the center to monitor issues such as reproducibility.
- Used a python based tool, which compiles, launches, and checks the outputs of parallel jobs. The tool is managed through a continuous integration of Jenkins server.
- Every application testing script or benchmark suite is made available as an elementary test unit. Regularly, through well-defined campaigns or random pick-up, a set of these tests will be initiated and run on Shaheen II via Jenkins.
- The results (performance or accuracy) will be carefully archived by Jenkins along with the detailed system state on the service, login or computer nodes at that time.

Shaheen II Regression Workflow Complete cycle



Jenkins Script

Build History (trend)

#3193	Apr 29, 2015 8:45:21 AM
#3192	Apr 29, 2015 8:41:21 AM
#3191	Apr 29, 2015 8:37:21 AM
#3190	Apr 29, 2015 8:33:21 AM
#3189	Apr 29, 2015 8:29:21 AM
#3188	Apr 29, 2015 8:25:21 AM
#3187	Apr 29, 2015 8:21:21 AM
#3186	Apr 29, 2015 8:17:21 AM
#3185	Apr 29, 2015 8:13:21 AM
#3184	Apr 29, 2015 8:09:21 AM
#3183	Apr 29, 2015 8:05:21 AM
#3182	Apr 29, 2015 8:01:21 AM
#3181	Apr 29, 2015 7:57:21 AM
#3180	Apr 29, 2015 7:53:21 AM
#3179	Apr 29, 2015 7:49:21 AM
#3178	Apr 29, 2015 7:45:21 AM
#3177	Apr 29, 2015 7:41:21 AM
#3176	Apr 29, 2015 7:37:21 AM
#3175	Apr 29, 2015 7:33:21 AM
#3174	Apr 29, 2015 7:29:21 AM
#3173	Apr 29, 2015 7:25:21 AM
#3172	Apr 29, 2015 7:21:21 AM
#3171	Apr 29, 2015 7:17:21 AM
#3170	Apr 29, 2015 7:13:21 AM
#3169	Apr 29, 2015 7:09:21 AM
#3168	Apr 29, 2015 7:05:21 AM
#3167	Apr 29, 2015 7:01:21 AM
#3166	Apr 29, 2015 6:57:21 AM
#3165	Apr 29, 2015 6:53:21 AM
#3164	Apr 29, 2015 6:49:21 AM

More ...

RSS for all RSS for failures

Max # of builds to keep
if not empty, build records are only kept up to this number of days

if not empty, only up to this number of build records are kept

- [ ] This build is parameterized
- [x] Disable Build (No new builds will be executed until the project is re-enabled.)
- [x] Execute concurrent builds if necessary (beta)
- [x] Restrict where this project can be run

Label Expression: shaheen_2

Advanced Project Options

Source Code Management

- (•) None

Build Triggers

- [ ] Build after other projects are built
- [ ] Trigger builds remotely (e.g., from scripts)
- [x] Build periodically

Schedule: */4 * * * *

- [ ] Poll SCM

Build

Execute shell

Command:

```
export WORK=$HOME/tmp/$USER/jenkins-${JOB_NAME}-${BUILD_NUMBER}

mkdir -p  $WORK

cd $WORK

tar fcv zephyr.tar -C $HOME ZEPHYR > tar_fcv.out 2> tar_fcv_err

if [ $? -ne 0 ] ; then
        echo "FAILED : could not create the tar"
        echo ----- out ---------------
        cat tar_fcv.out
        echo ----- err ---------------
        cat tar_fcv.err
        echo ----- end ---------------
```

Jenkins GUI

Jenkins

search log in

Jenkins » shaheen 2

ENABLE AUTO REFRESH

Build History

All | SAT | build | neser | osprey | **shaheen 2**

S	W	Name ↓	Last Success	Last Failure	Last Duration
		shaheen 2 filesystem	1 min 19 sec (#2950)	6 days 7 hr (#680)	20 sec
		shaheen 2 get powercap	33 min (#28)	N/A	0.21 sec
		shaheen 2 slurm is alive	1 min 19 sec (#4704)	11 hr (#4362)	15 sec
		shaheen 2 SPECFM	3 days 3 hr (#18)	3 days 5 hr (#9)	63 ms
		shaheen 2 ssh	3 min 19 sec (#567)	1 day 17 hr (#75)	0.59 sec
		shaheen 2 workload	19 sec (#2344)	7 hr 35 min (#2253)	7.1 sec
		shaheen 2 xtnodestat	3 min 19 sec (#926)	N/A	2.1 sec

Icon: S M L

Legend RSS for all RSS for failures RSS for just latest builds

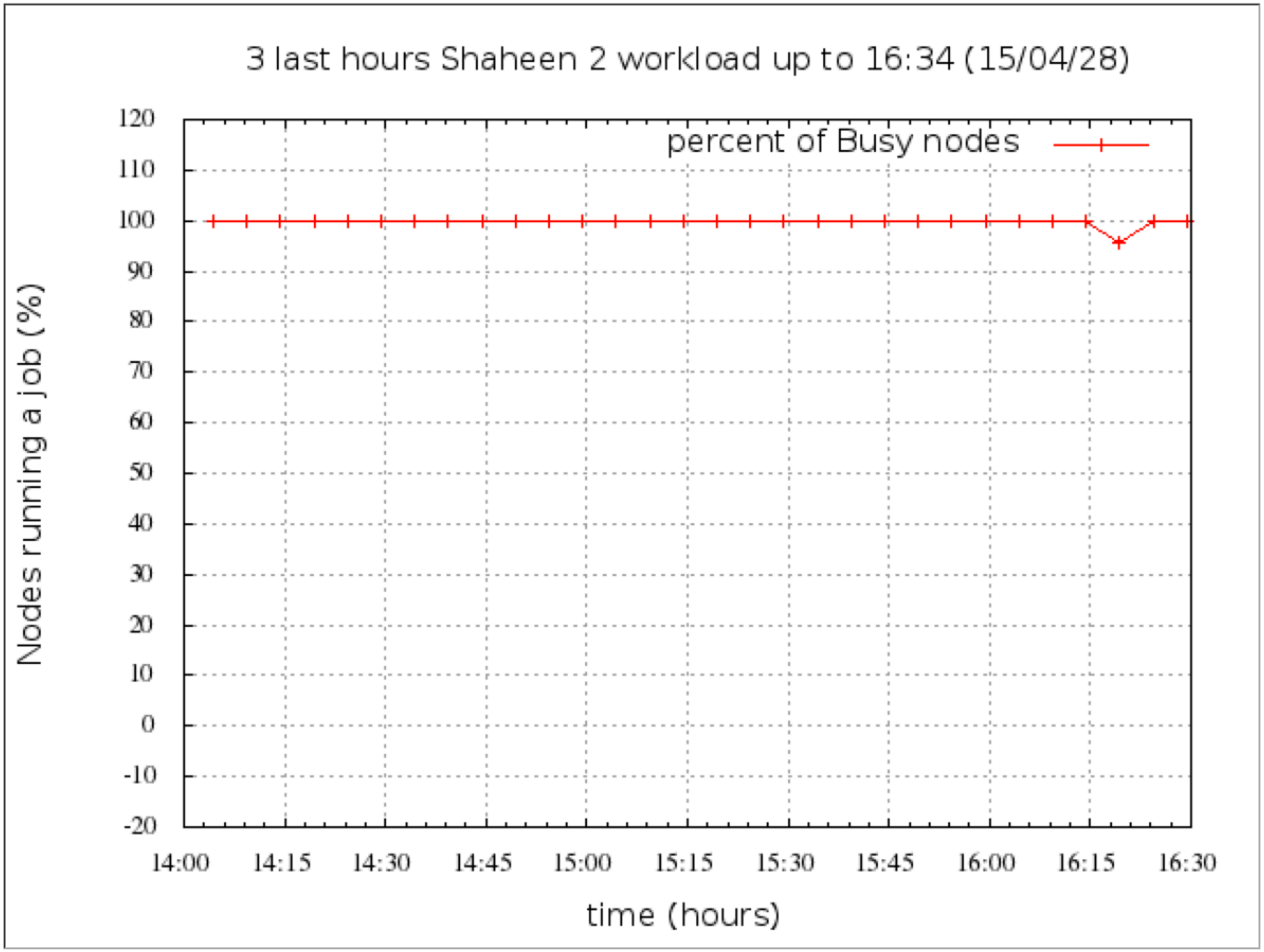
Page generated: Apr 28, 2015 4:34:41 PM Jenkins ver. 1.424.6

Workload monitor

log **non-log**

1 h
3 h
6 h
9 h
12 h
18 h
24 h

config : **6174**
resv : **6144**
use : **6144**
avail : **17**
down : **13**
load : **99 %**



xtnodestats

Load Peaks

99% 95%

Jobs OK

2015-04-28_16-29-21
2015-04-28_16-24-21
2015-04-28_16-19-21

Jobs Failed

None

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

Conclusions

- With acquisition of the new Cray XC40, Shaheen II, KAUST is once again the owner of a world-class supercomputer.

- Shaheen II will enable and grow collaboration with several in-Kingdom universities, industrial partners and other international leadership class supercomputers centers.

- XALT+DARSHAN+SLURM+SWTool (all open-source)
  → Real image of what is happening in the system
  → Preventive detection of performance issue

- Shaheen II software ecosystem : Jenkins will allow to detect any regression


Thanks !
© Shadi Khoja Photography 2012