Implementation of grid Tier 2 and Tier 3 facilities on a Distributed OpenStack Cloud

Speaker : Antonio Limosani

On Behalf of CoEPP Cloud Team :

Lucien Boland, Paul Coddington. Sean Crosby, **Joanna Huang**, **Martin Sevior**, Ross Wilson, Shunde Zhang

University of Adelaide

University of Melbourne

University of Sydney

eResearch South Australia

rc-tech@coepp.org.au

Outline

- Our Research Centre
- NeCTAR Research Cloud
- Project Overview
- Tier 2 and Tier3 Clouds
- Summary

ARC CoEPP (Client)

Australian Research Council funded "Centre of Excellence of Particle Physics at the TeraScale"

Brings together experimentalists and theorists to focus on the era of discovery launched by the LHC .e.g. Study of the Higgs Boson, Searching for Supersymmetry and other models of new physics

Began life in 2011

THE UNIVERSITY OF MELBOURNE

THE UNIVERSITY of ADELAIDE

Altogether ~80 people incl. academics, post-docs and post-grads

THE UNIVERSITY OF SYDNEY

MONASH University

Partner Institutes

Duke UNIVERSITY

UNIVERSITÉ DE GENÈVE

UNI FREIBURG

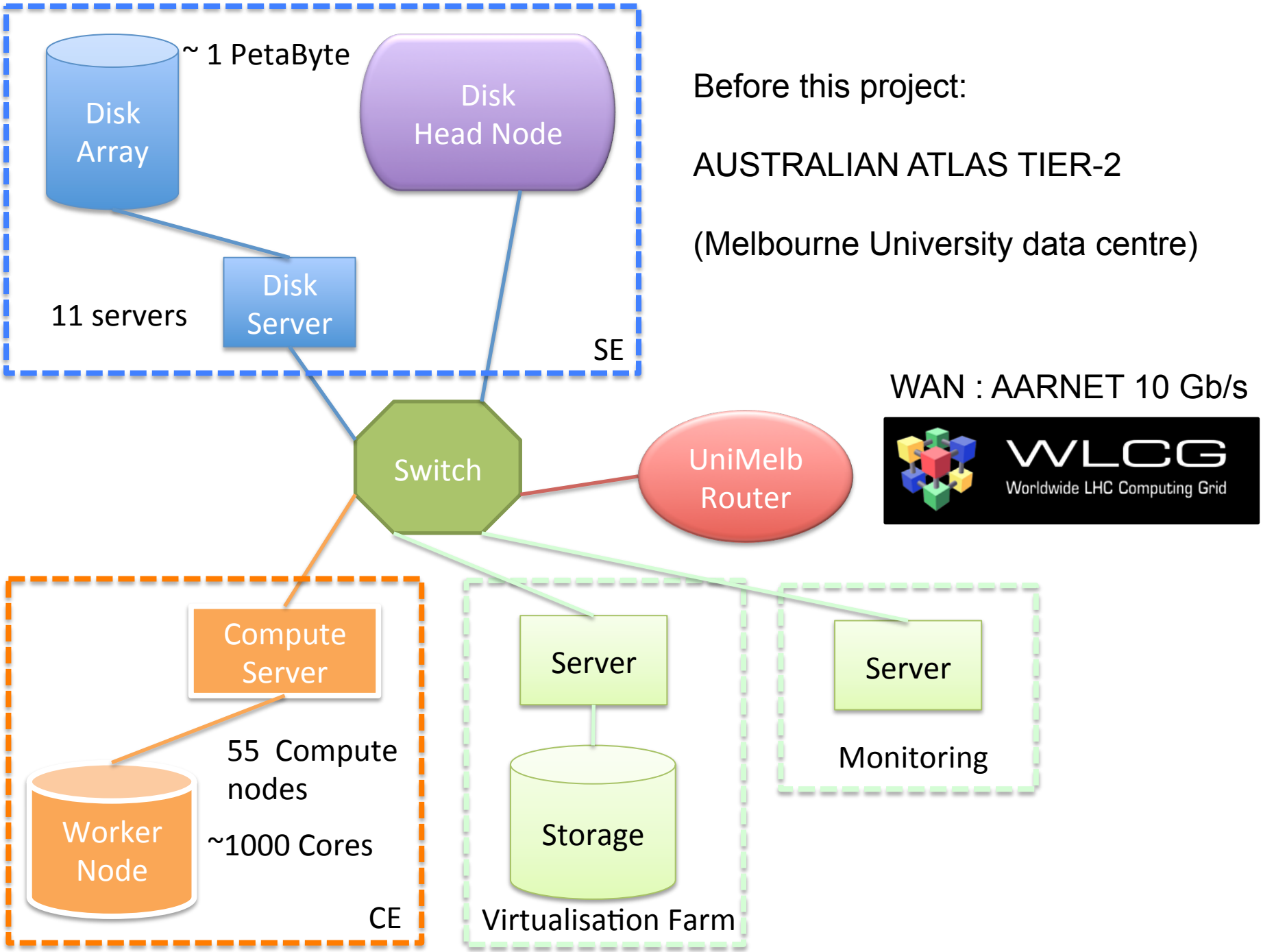
UNIVERSITY OF CAMBRIDGE

INFN Istituto Nazionale di Fisica Nucleare

Penn UNIVERSITY of PENNSYLVANIA

CoEPP ARC Centre of Excellence for Particle Physics at the Terascale

nectar



Before this project:

AUSTRALIAN ATLAS TIER-2

(Melbourne University data centre)

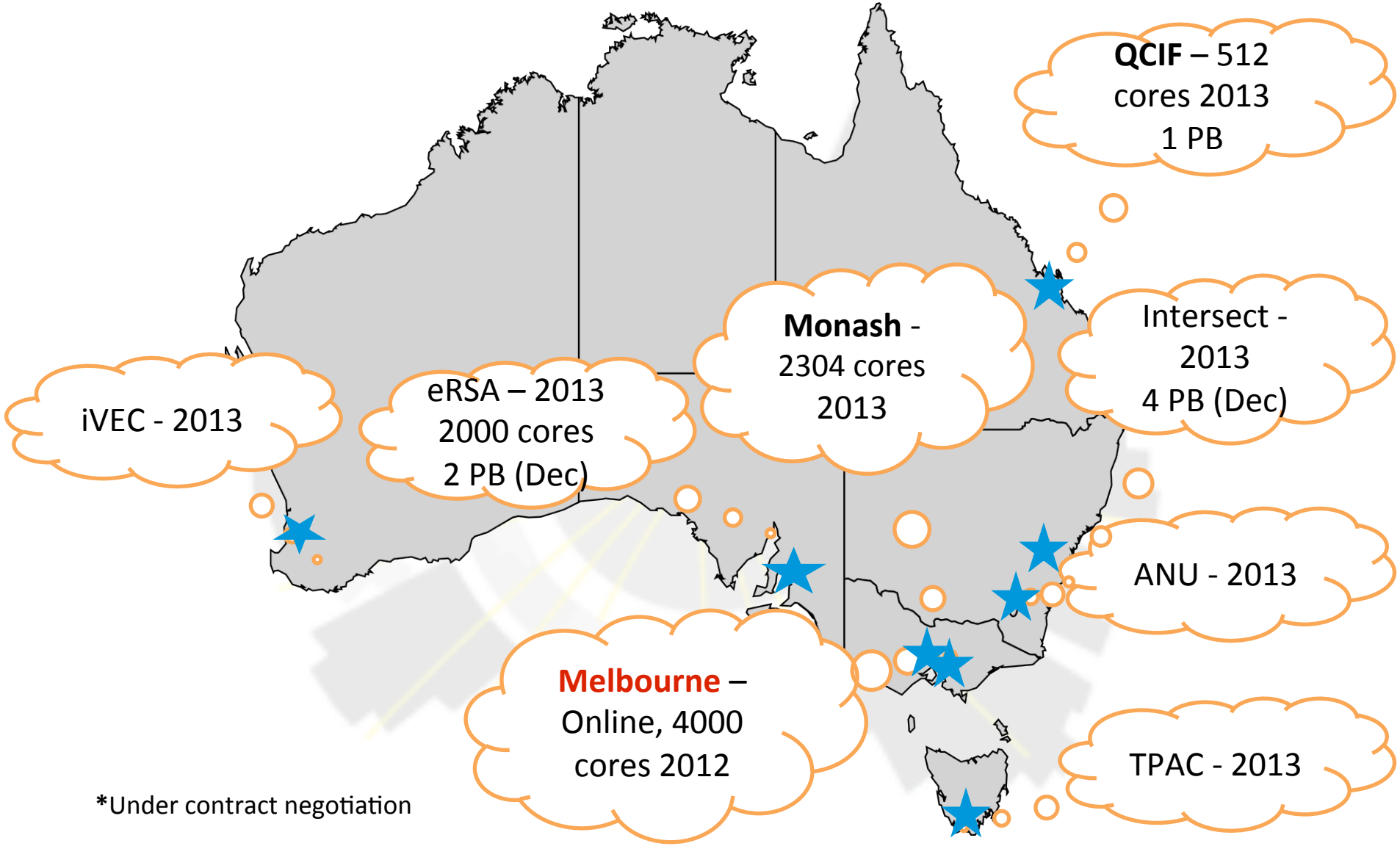
NeCTAR Research Cloud

- NeCTAR (*National eResearch Collaboration Tools and Resources*)
  - is a $47M Australian government funded project aiming to build new infrastructure specifically for the needs of Australian researchers
- NeCTAR Research Cloud
  - One of NeCTAR's four programs is the creation of a 30,000 cores Infrastructure-as-a-service cloud spanning 8 locations around Australia
  - Based on OpenStack (now on Grizzly)

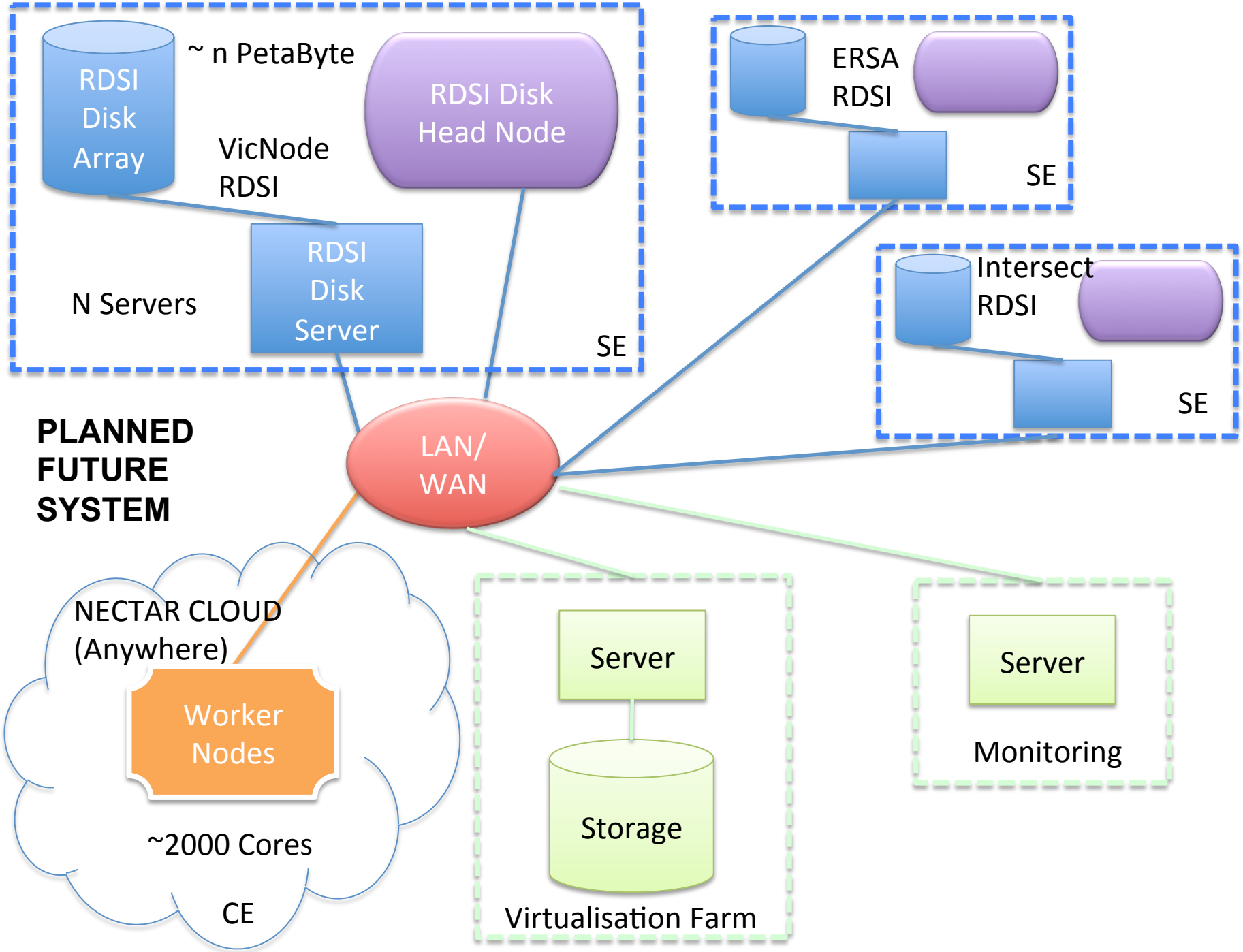
openstack

- RDSI (Research Data Storage Infrastructure) - $50M Australian government funded project. 20 PB

From One Node ... to Eight
From 4,000 Cores ... to 30,000 in 2013-2014



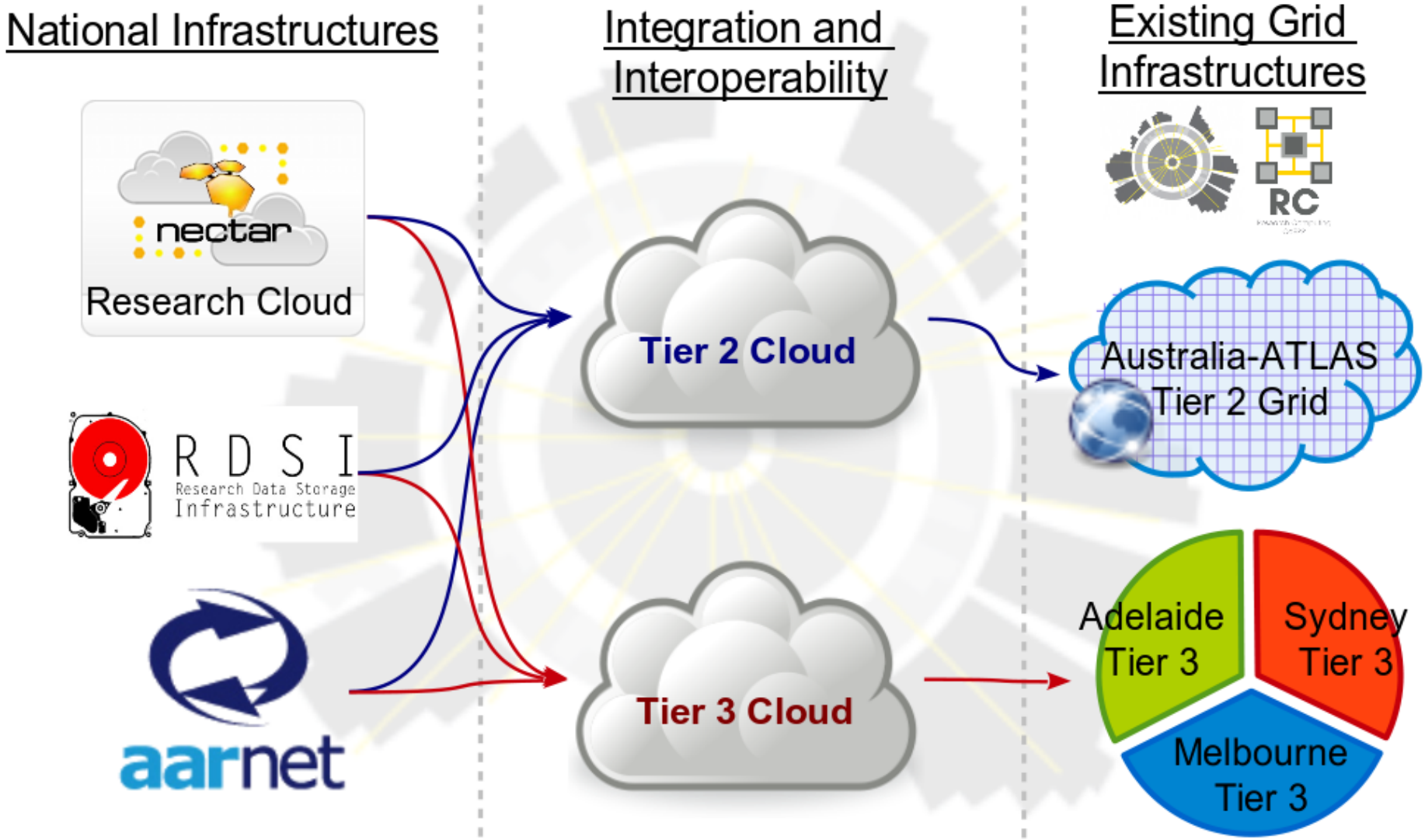
*Under contract negotiation


RDSI Disk Array
~ n PetaByte
RDSI Disk Head Node
VicNode RDSI
RDSI Disk Server
N Servers
SE
ERSA RDSI
SE
Intersect RDSI
SE
PLANNED FUTURE SYSTEM
LAN/ WAN
NECTAR CLOUD (Anywhere)
Worker Nodes
~2000 Cores
CE
Server
Storage
Virtualisation Farm
Server
Monitoring

High Throughput Computing for Globally Connected Science

- This project was funded under NeCTAR's eResearch Tools program – aiming to fix research capability gaps
- 1.5 years duration, 2 sub-projects
  - To augment Australian ATLAS Tier 2 capacity
  - To build a federated Tier 3 for high throughput data analysis and processing
- CoEPP Research Computing Centre will take over operations and maintenance after the end of 2013

Project Scope

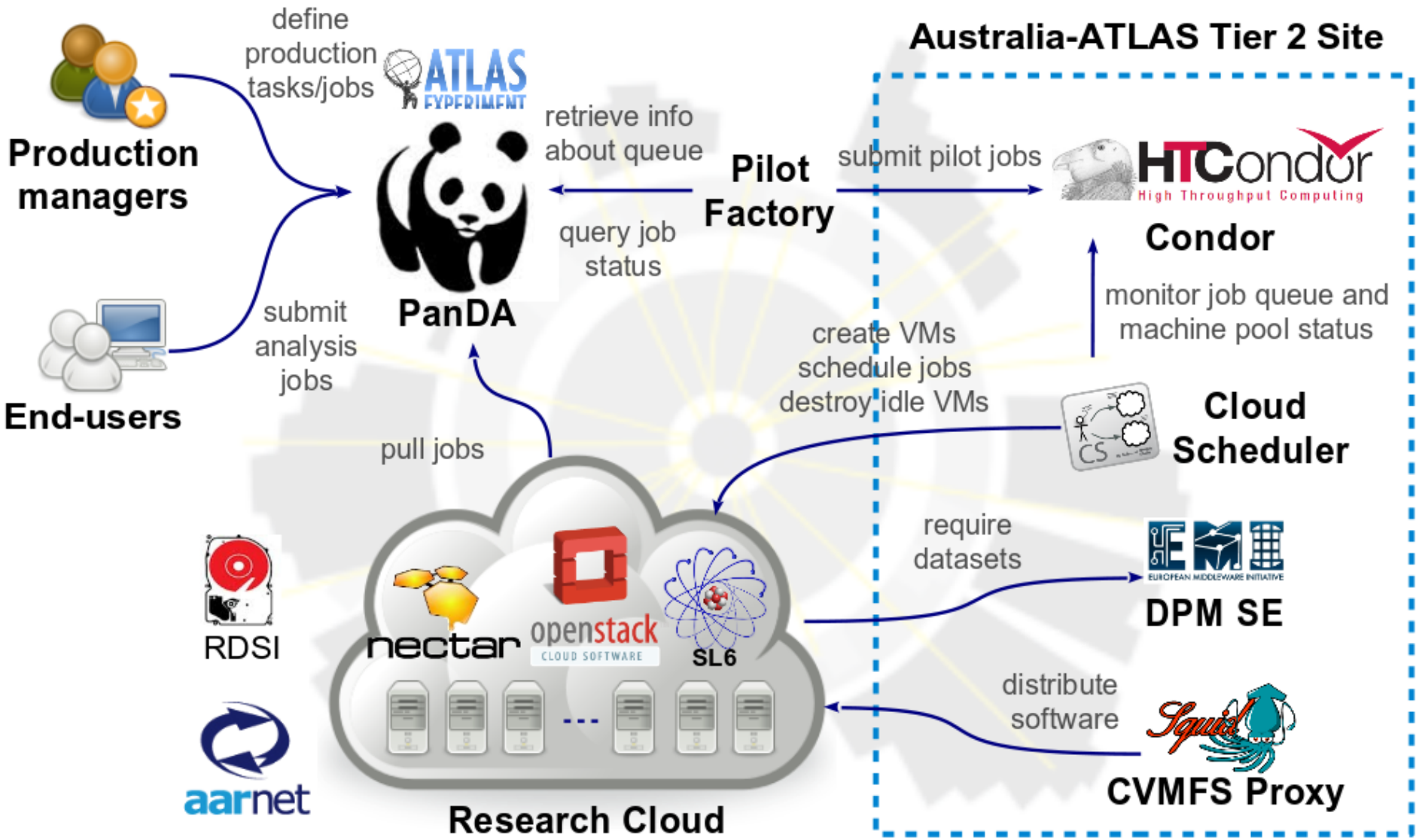


Tier 2 Cloud – Overview

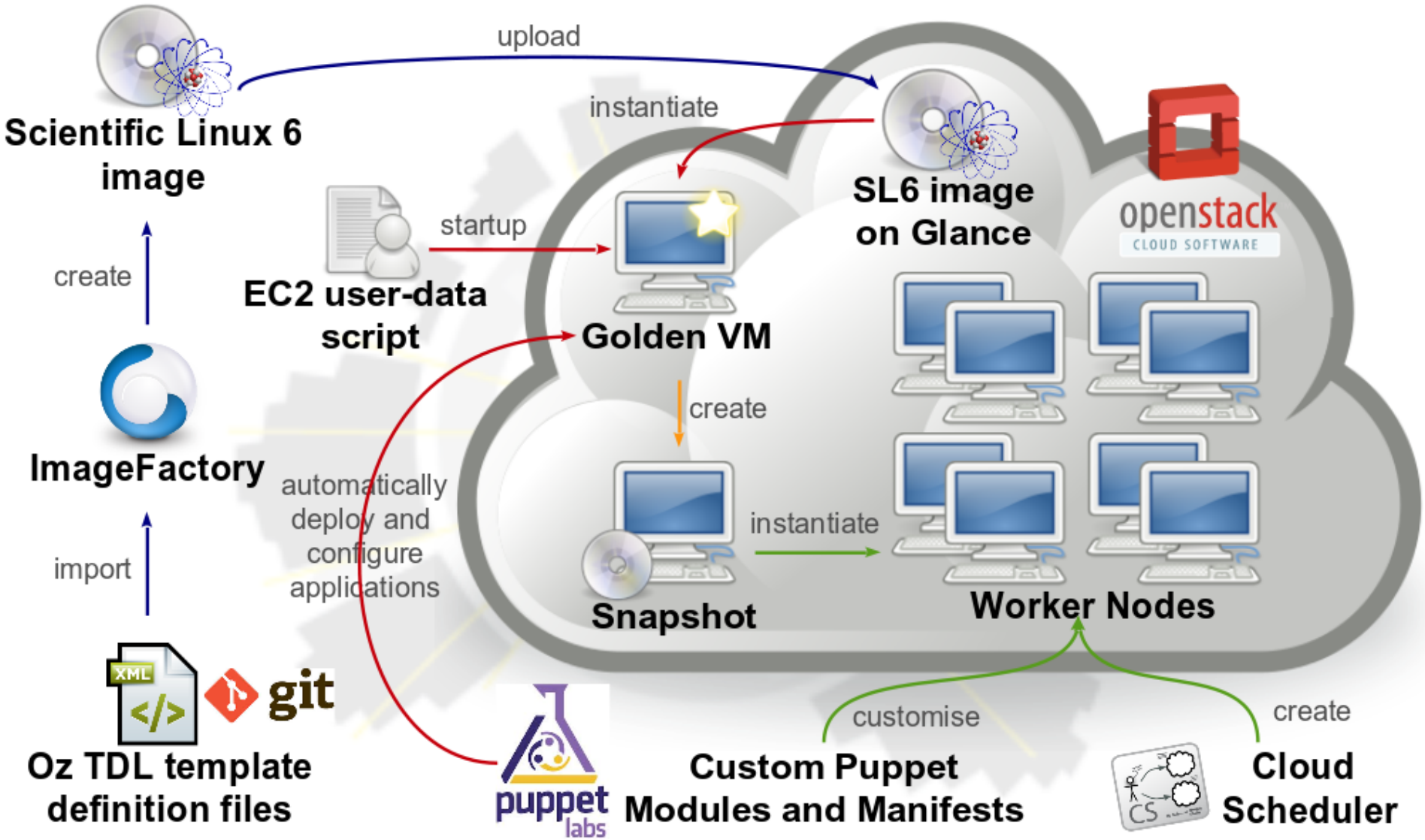
- Integration with ATLAS PanDA job framework
- Interoperability with OpenStack cloud
- Dynamic scalability using Condor batch system and Cloud Scheduler software (U. Victoria, Canada)
- CVMFS server to distribute software libraries
- Automated service configuration via Puppet
- Computing capacity of 2000 cores
- Storage capacity of 2PB

… to be able to run ATLAS MC production jobs and user analysis jobs on NeCTAR's OpenStack cloud

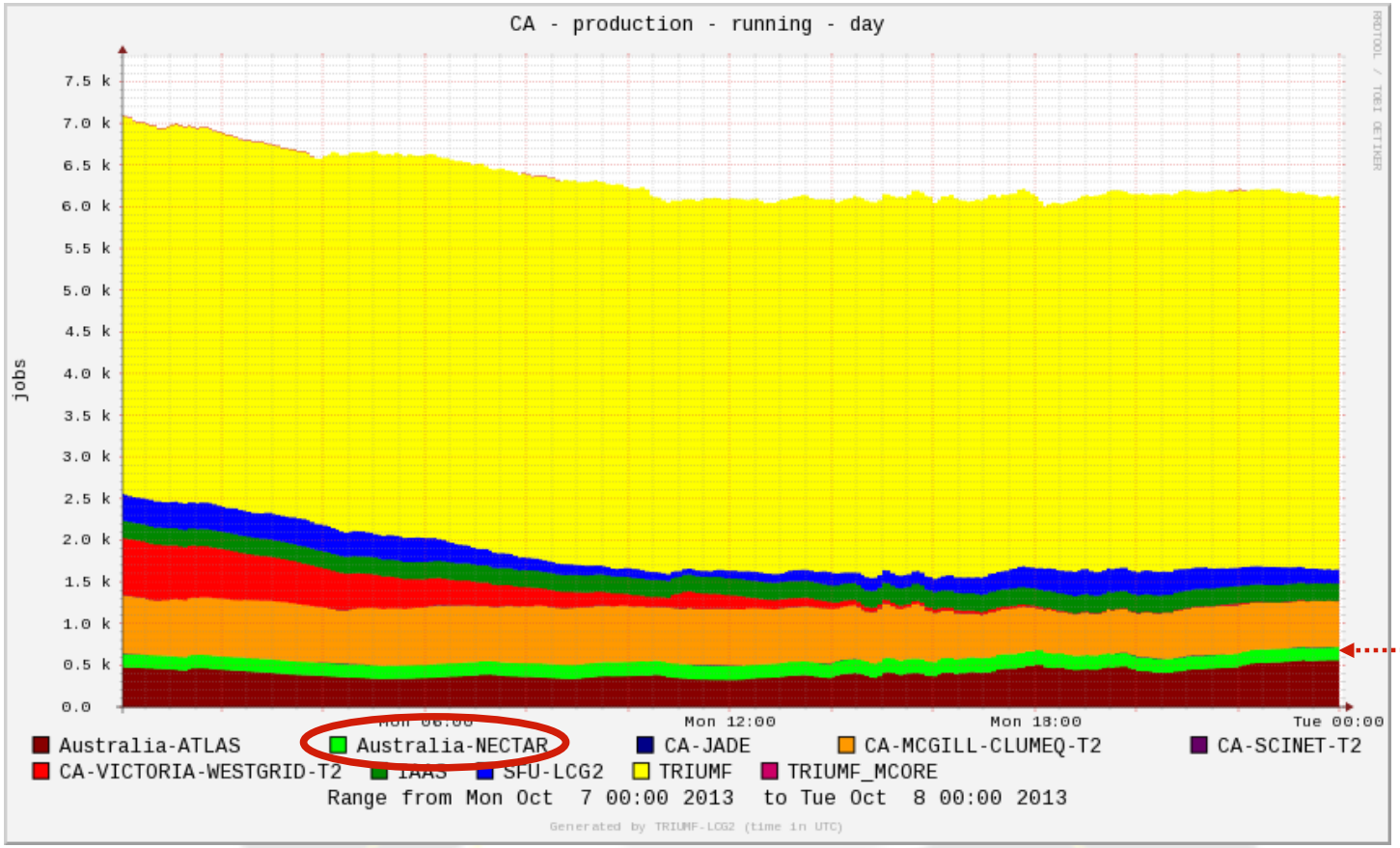
Tier 2 Cloud – System Framework



Virtual Machine Contextualisation



Tier 2 Cloud – Production Jobs



NeCTAR

- 160 cores running ATLAS MC production jobs
  - 20 8-cores SL6 machines on NeCTAR cloud

Tier 2 Cloud – Future Work Plans

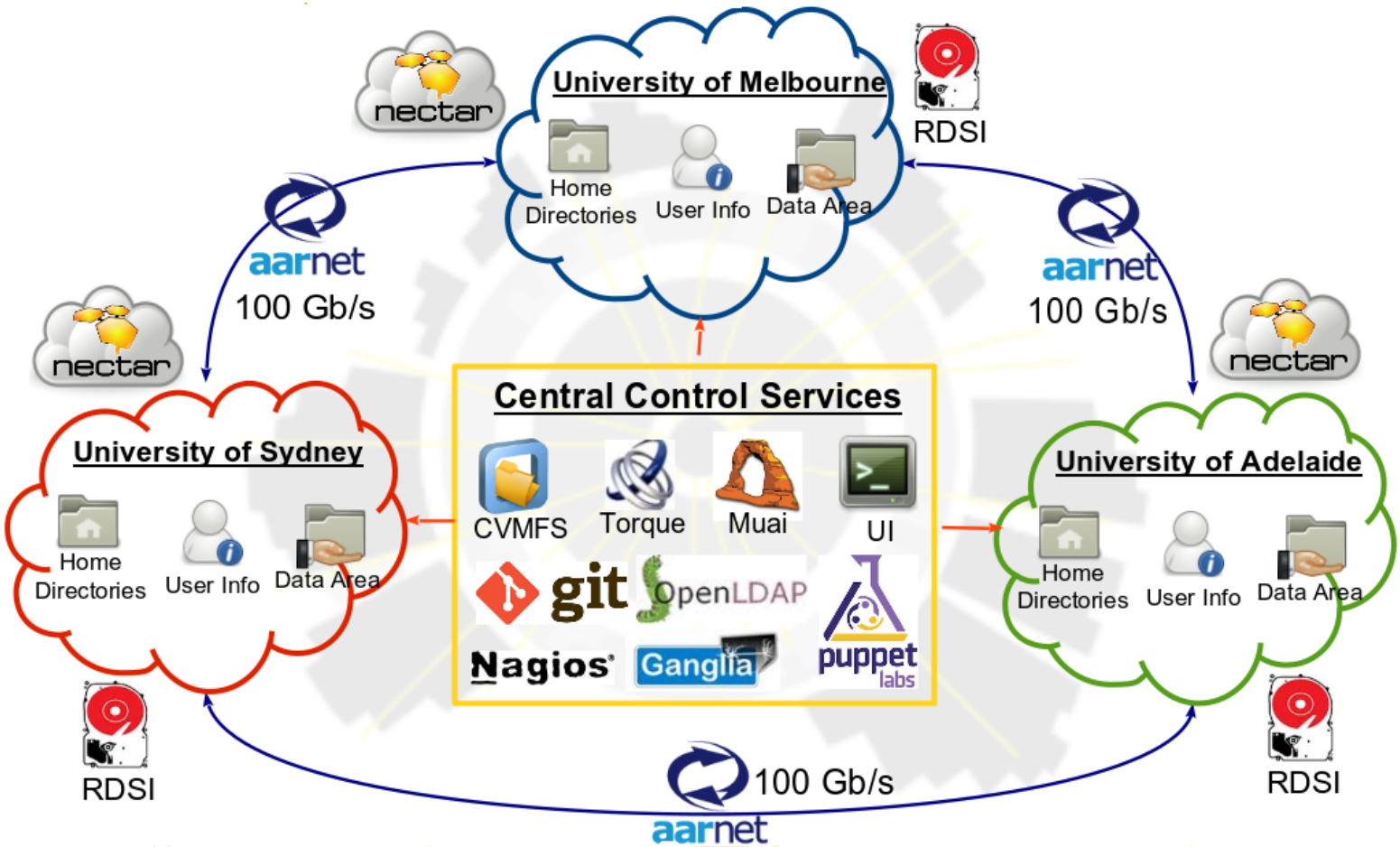
- Ensuring high system robustness and reliability
- Tuning performance for ATLAS analysis jobs
  - e.g. filesystem, network, I/O
- Connecting Tier 2 cloud to Belle II infrastructure
- Expanding computer and storage capacity
- Dealing with potential data locality issue
  - Unclear implementation details, e.g. data presenting interfaces and filesystem architecture
  - Shared 100Gb/s network connectivity

Tier 3 Cloud – Overview

- Integration of OpenStack cloud facilities and local resources at each CoEPP node of Melbourne, Adelaide, Sydney, and Monash
- Torque system function extended by further development for dynamic resource allocation and provisioning
- CVMFS server to host and distribute software and libraries to meet local users' software requirement
- Identical access to files in User Home Directories from any CoEPP node
- Large Data Area to store data sets commonly required or outputs produced by data analysis, simulation and processing work
- Multi-core high performance Interactive Nodes

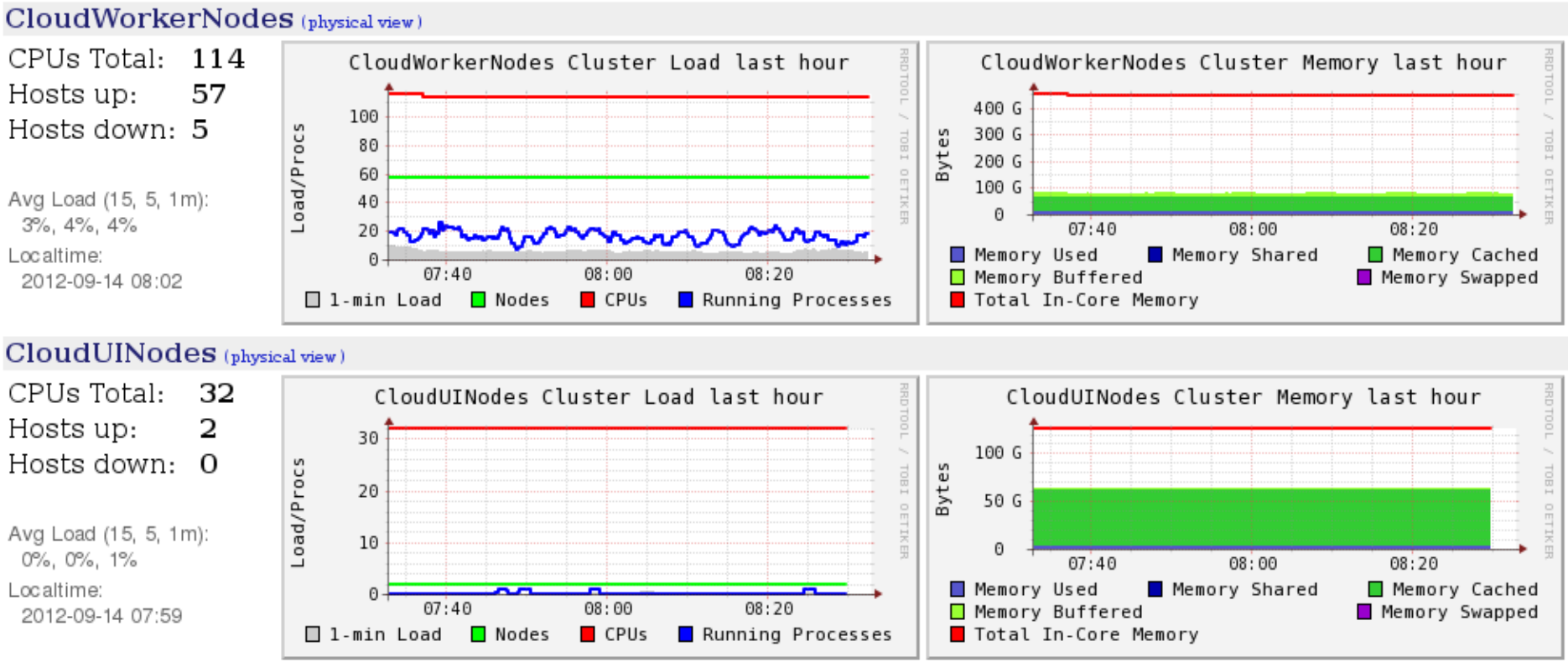
… to be able to run large high throughput data analysis, processing and simulation jobs across all CoEPP nodes

Tier 3 Cloud – System Framework



- LDAP+Kerberos: Authentication and authorisation architecture
- NFS, Ceph, OrangeFS, etc.: Distributed users' home directories (Anycast)
- XrootdFS: Distributed data area and single data access entrance

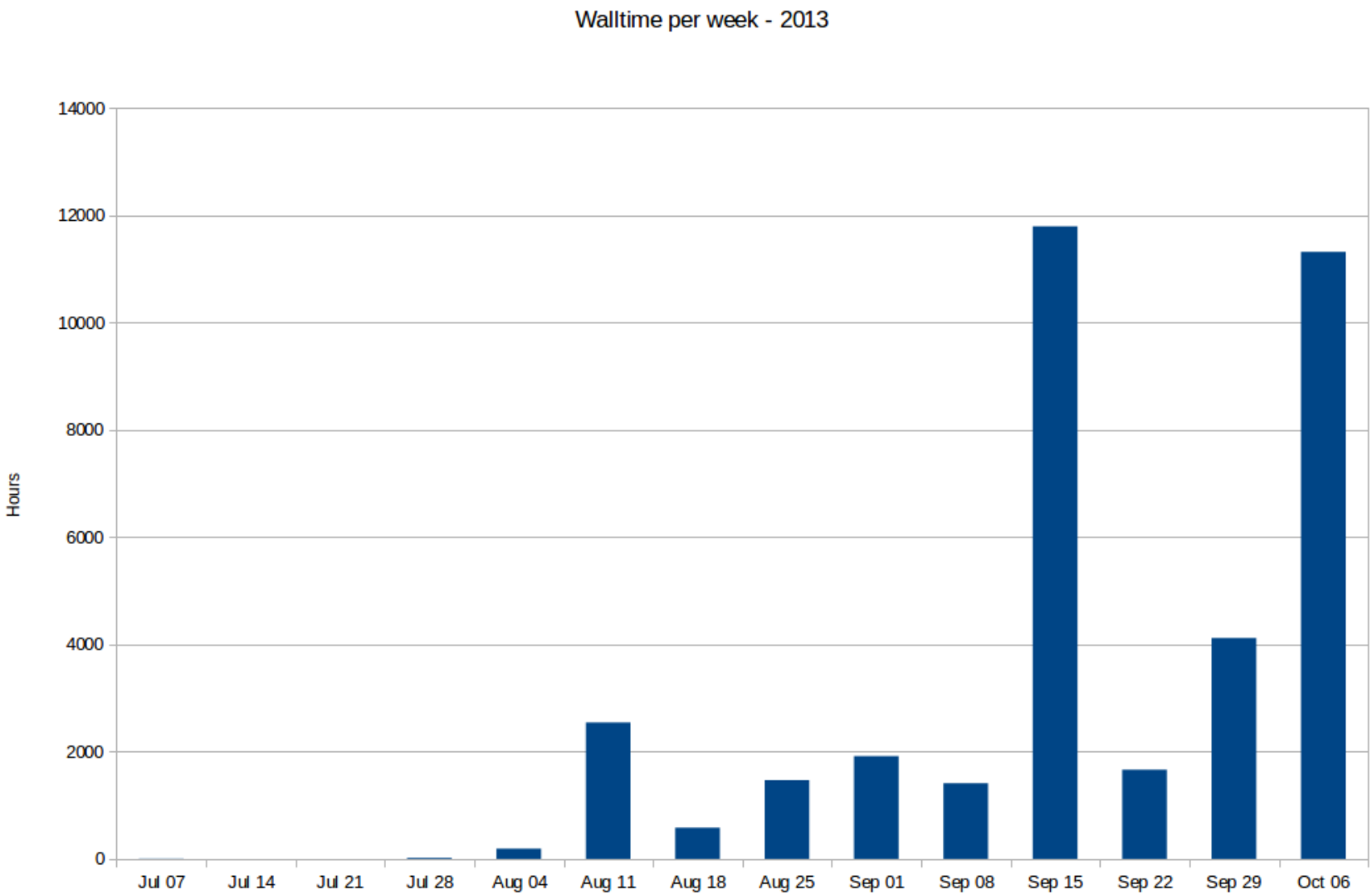
Tier 3 Cloud in Production



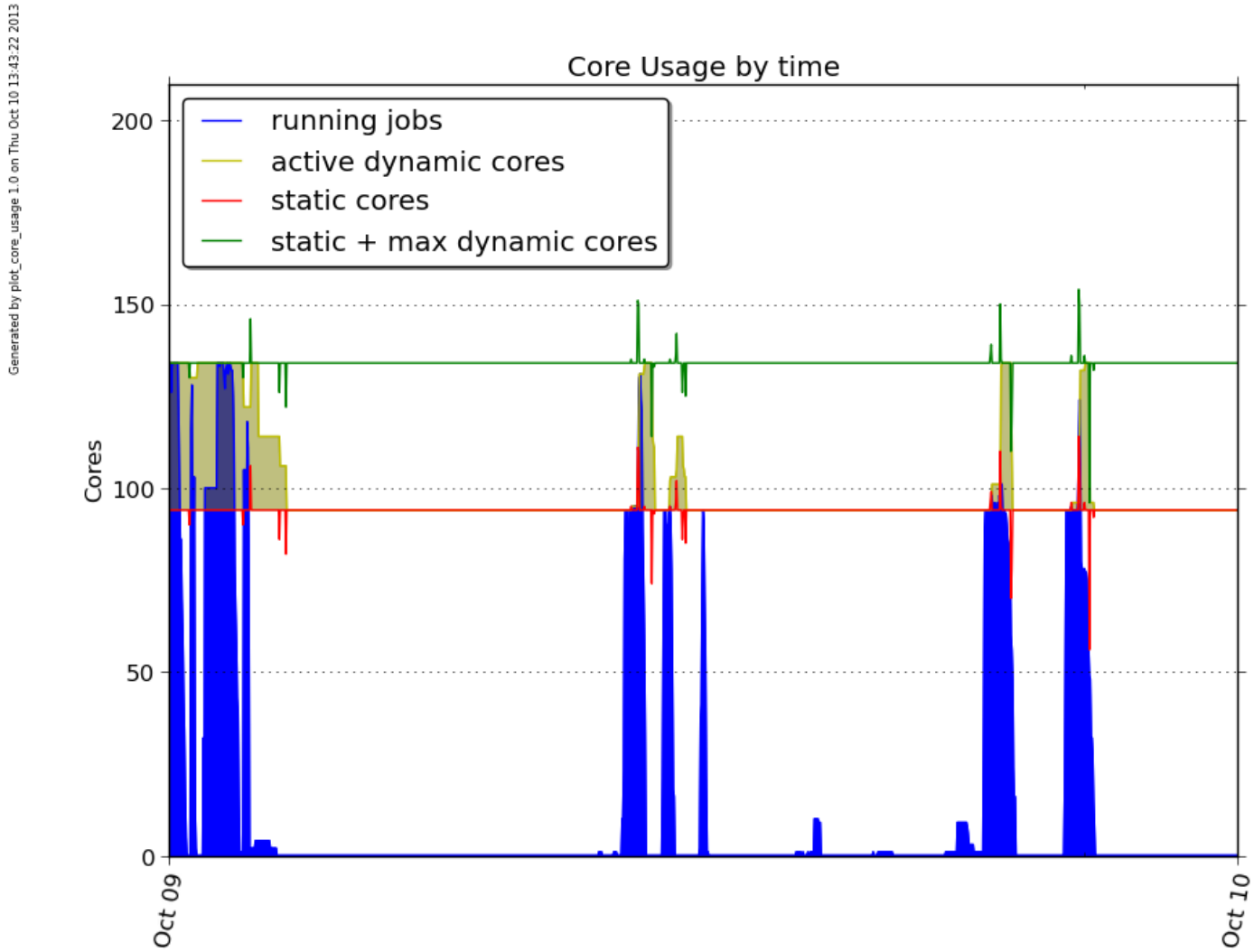
Using central control services we can build a new Tier 3 node on the cloud within minutes

- 90 static + 40 dynamic single-core SL6 worker nodes
- 2 16-cores interactive nodes with 480GB local storage

Cloud Cluster: CPU (Wall time) Hours per Week



Cloud Cluster: Dynamic Torque in action



Shunde Zhang of eRSA
extended torque through VM pool to create additional VMs for the batch system on a needs basis

"Dynamic Torque" - cloud-based dynamic VM provisioning

See Lucien Boland's presentation. "
Dynamic VM provisioning for Torque in a cloud environment"
Tuesday 15:45

Additional Virtual Organizations

- Dynamic Torque can be interfaced with the CREAM CE in the Australian ATLAS Tier2 Site
- CVMFS server to provide access to software builds of other VOs e.g. Belle II
- Australian Tier2 will soon be able to support the Belle II VO
- Propagate these tools to other research communities e.g. LIGO Collaboration in Australia

Tier 3 Cloud – Future Work Plans

- Duplicate current Tier 3 cloud resources in Melbourne to other CoEPP nodes
- Research distributed file system solution for hosting users' home directories
- Deploy Xrootd(FS) for distributing shared data area
  - Collaborating with Doug Benjamin (Duke U)
  - File Residency Manager (FRM) caching
- Integrate our distributed file system with external distributed storage provided by RDSI
- Consolidate and manage an optimized cloud-based infrastructure

Summary

- Australian Tier 2 cloud is now fully functioning with operations to run large-scale production MC simulation for ATLAS
- CoEPP federated Tier 3 cloud is now in production providing extra computing resources on cloud to all collaborators of CoEPP
- Continuing to improve and manage our cloud infrastructure in a reliable and automated fashion, as well as simplify operational process for reducing maintenance cost
- Current infrastructure is cloud-ready and allows for future growth and expansion via further consolidation

Career opportunities

http://www.coepp.org.au/career-opportunities

RESEARCH COMPUTING SYSTEMS ADMINISTRATOR

COEPP
SCHOOL OF PHYSICS
REFERENCE NO. 1950/0913

- **Large Linux installation site administration including remote system deployment, maintenance and monitoring**
- **Work with the next generation of cluster computing technologies and contribute to enabling cutting-edge particle physics research.**
- **Full-time fixed-term, remuneration package: $92K p.a. which includes leave loading and 17% super**

THE UNIVERSITY OF SYDNEY

Extra Slides

Tier 3 Cloud – Central Services

- Federation-wide central control services
  - Puppet: Configuration management and automated deployment
  - CVMFS: Experiment software repositories
  - LDAP+Kerberos: Authentication and authorisation architecture
  - Torque+Maui+extension: Dynamic computer cluster manager
  - NFS, Ceph, OrangeFS, etc.: Distributed users' home directories (under evaluation)
  - XrootdFS: Distributed data area and single data access entrance
  - Nagios+Ganglia: Service availability and performance monitoring
  - UI: Batch job submission node with anycast methodology implemented

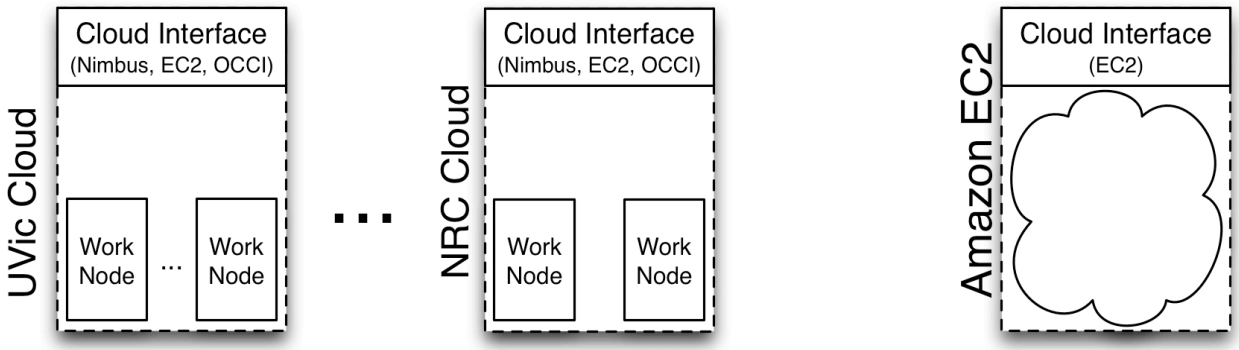
… to be able to utilise central services to deploy, manage, monitor Tier 3 clouds on NeCTAR's OpenStack cloud

Federated Xrootd storage clusters

- In development with Doug Benjamin (Duke University – COEPP partner) and member of **Xrootd collaboration**
- Attracted to xrootd due to ease of implementation and flexibility e.g. authentication (non ATLAS members)
- Provide common namespace and visible directory tree in storage clusters spread across Australia
- Allows clients to easily identify, transfer and process data directly or via the batch system
- File residency manager (FRM) service allows for caching of data sets, and so will improve throughput across sites.

Extra Slide: Cloud Scheduler Workflow

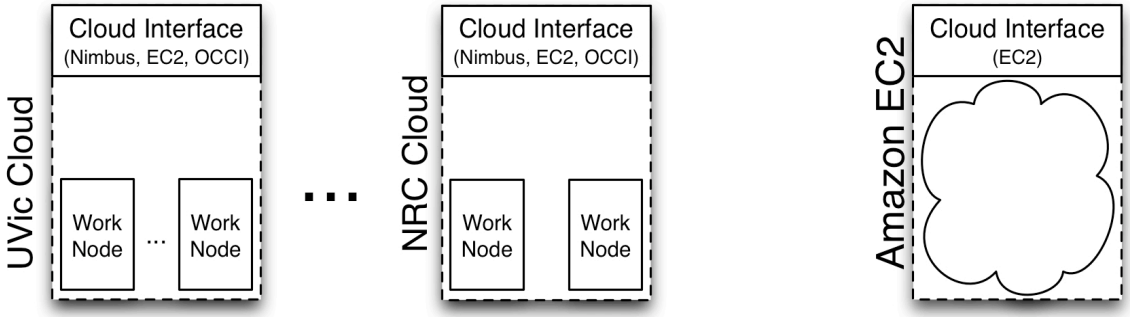
Step 1

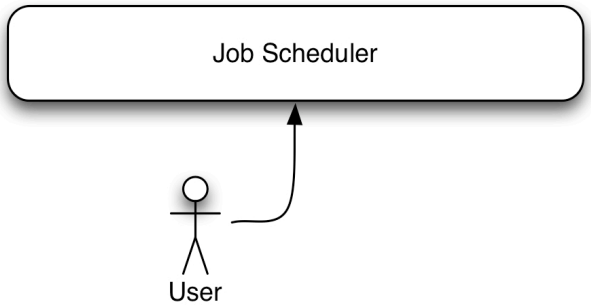


Research and Commercial clouds made available with some cloud-like interface.

Extra Slide: Cloud Scheduler Workflow

Step 2

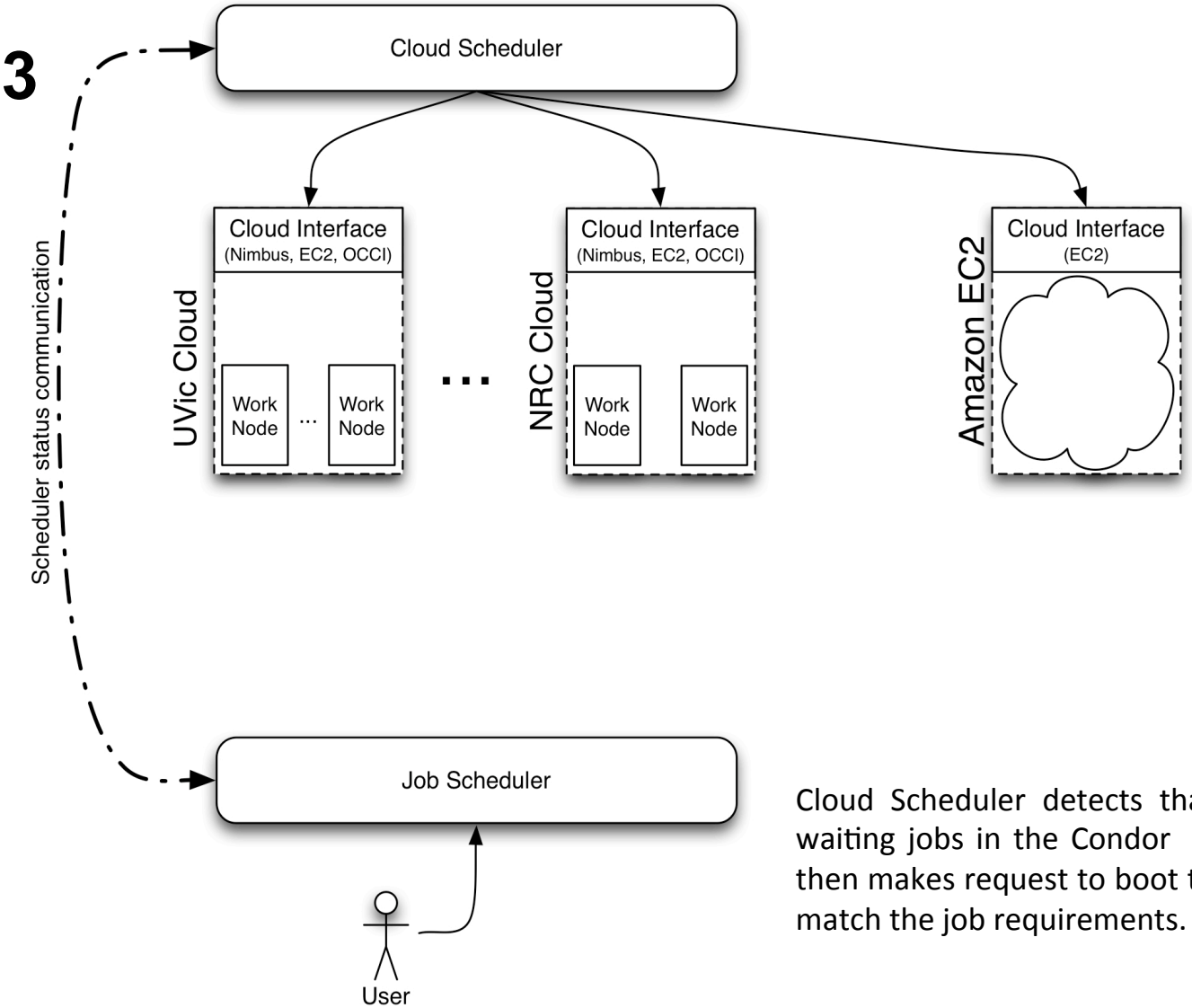




User submits to Condor Job scheduler that has no resources attached to it.

Extra Slide: Cloud Scheduler Workflow

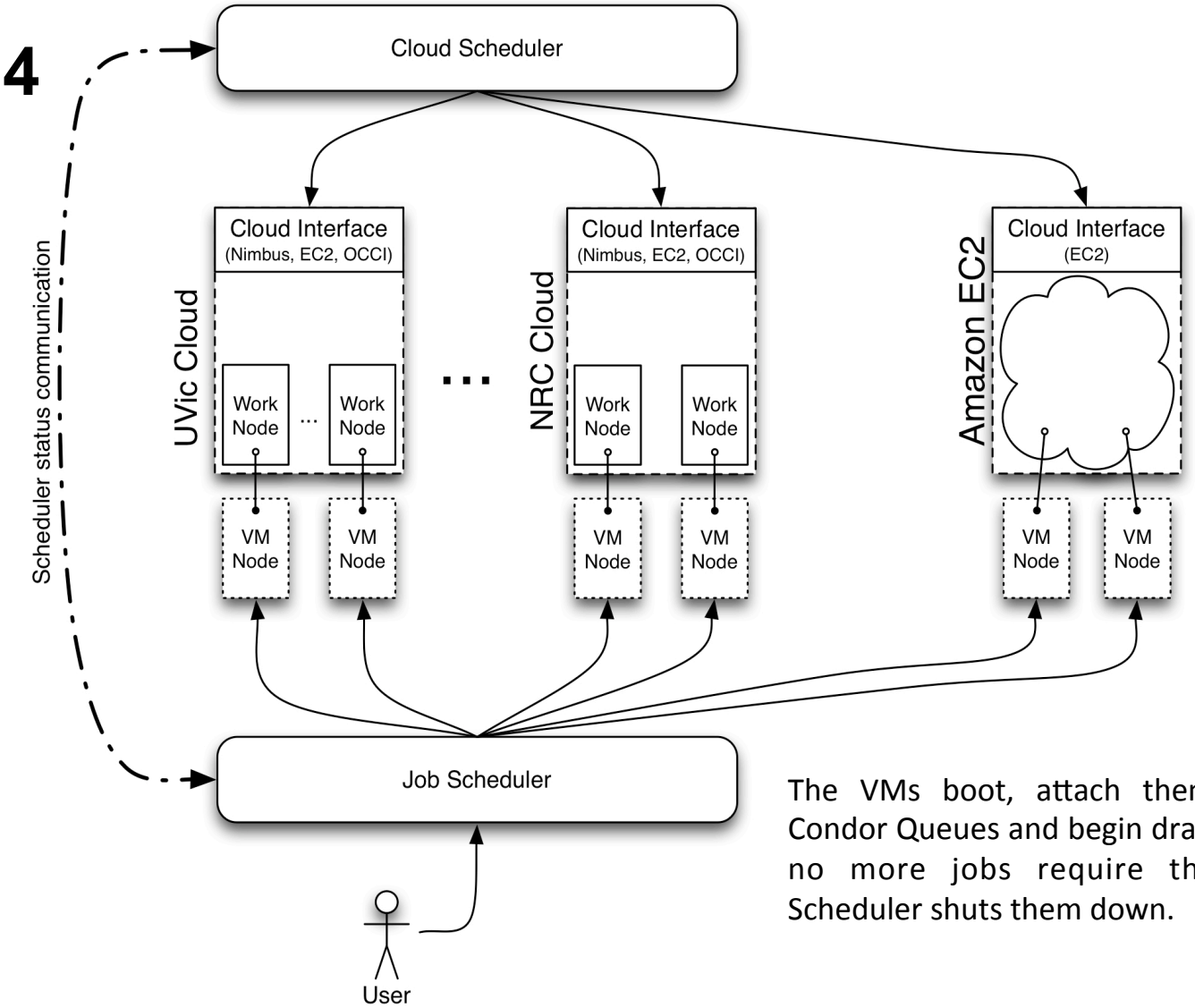
Step 3



Cloud Scheduler detects that there are waiting jobs in the Condor Queues and then makes request to boot the VMs that match the job requirements.

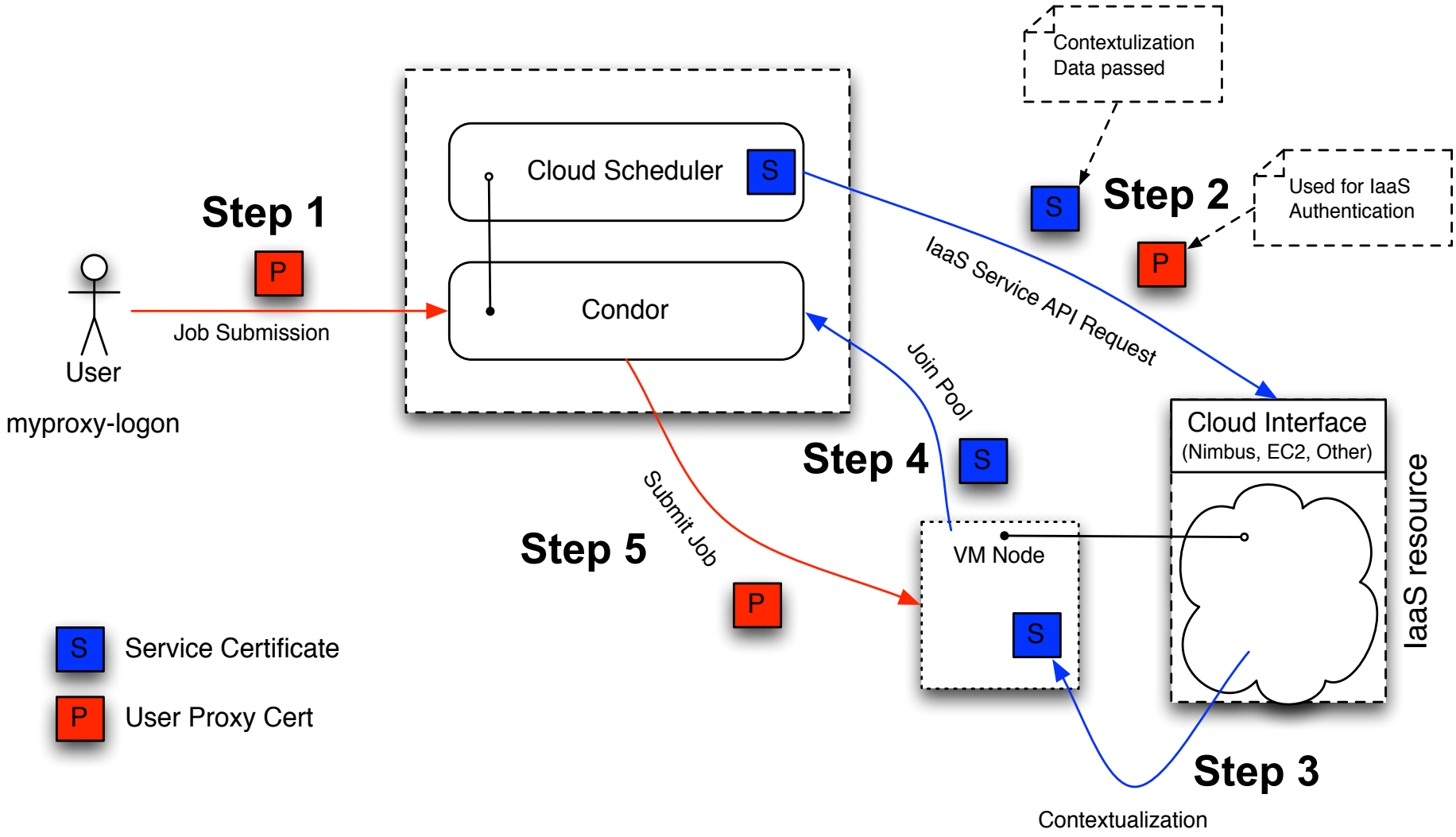
Extra Slide: Cloud Scheduler Workflow

Step 4



The VMs boot, attach themselves to the Condor Queues and begin draining jobs. Once no more jobs require the VMs Cloud Scheduler shuts them down.

Extra Slide: CS Credential Transport



Glance

The Glance project provides services for discovering, registering, and retrieving virtual machine images. Glance has a RESTful API that allows querying of VM image metadata as well as retrieval of the actual image.

VM images made available through Glance can be stored in a variety of locations from simple filesystems to object-storage systems like the OpenStack Swift project.

Glance, as with all OpenStack projects, is written with the following design guidelines in mind:

Component based architecture: Quickly add new behaviors
Highly available: Scale to very serious workloads
Fault tolerant: Isolated processes avoid cascading failures
Recoverable: Failures should be easy to diagnose, debug, and rectify
Open standards: Be a reference implementation for a community-driven api

Ganglia

Ganglia is a scalable distributed monitoring system for high-performance computing systems such as clusters and Grids. It is based on a hierarchical design targeted at federations of clusters. It leverages widely used technologies such as XML for data representation, XDR for compact, portable data transport, and RRDtool for data storage and visualization. It uses carefully engineered data structures and algorithms to achieve very low per-node overheads and high concurrency. The implementation is robust, has been ported to an extensive set of operating systems and processor architectures, and is currently in use on thousands of clusters around the world. It has been used to link clusters across university campuses and around the world and can scale to handle clusters with 2000 nodes.

Puppet

Puppet is IT automation software that helps system administrators manage infrastructure throughout its lifecycle, from provisioning and configuration to orchestration and reporting. Using Puppet, you can easily automate repetitive tasks, quickly deploy critical applications, and proactively manage change, scaling from 10s of servers to 1000s, on-premise or in the cloud.

How Puppet Works

Puppet uses a declarative, model-based approach to IT automation.

Define the desired state of the infrastructure's configuration using Puppet's declarative configuration language.
Simulate configuration changes before enforcing them.
Enforce the deployed desired state automatically, correcting any configuration drift.
Report on the differences between actual and desired states and any changes made enforcing the desired state.

Nagios

Nagios is a powerful monitoring system that enables organizations to identify and resolve IT infrastructure problems before they affect critical business processes.

Designed with scalability and flexibility in mind, Nagios gives you the peace of mind that comes from knowing your organization's business processes won't be affected by unknown outages.

Nagios is a powerful tool that provides you with instant awareness of your organization's mission-critical IT infrastructure. Nagios allows you to detect and repair problems and mitigate future issues before they affect end-users and customers.

OZ TDL

TDL is an XML-based language for creating image templates. TDL files define aspects of a virtual machine image including operating system, installation settings, packages, and files. An Aeolus user imports TDL files into Conductor. Aeolus (specially Imagefactory and Oz) interprets these TDL files and builds them into images, which can be pushed into cloud providers.