Cloud Computing Activities in Australian ATLAS Tier-2/Tier-3

Joanna Huang, CoEPP

ATLAS Distributed Computing Tier-1/Tier-2/Tier-3 Jamboree

10th December 2012

Agenda

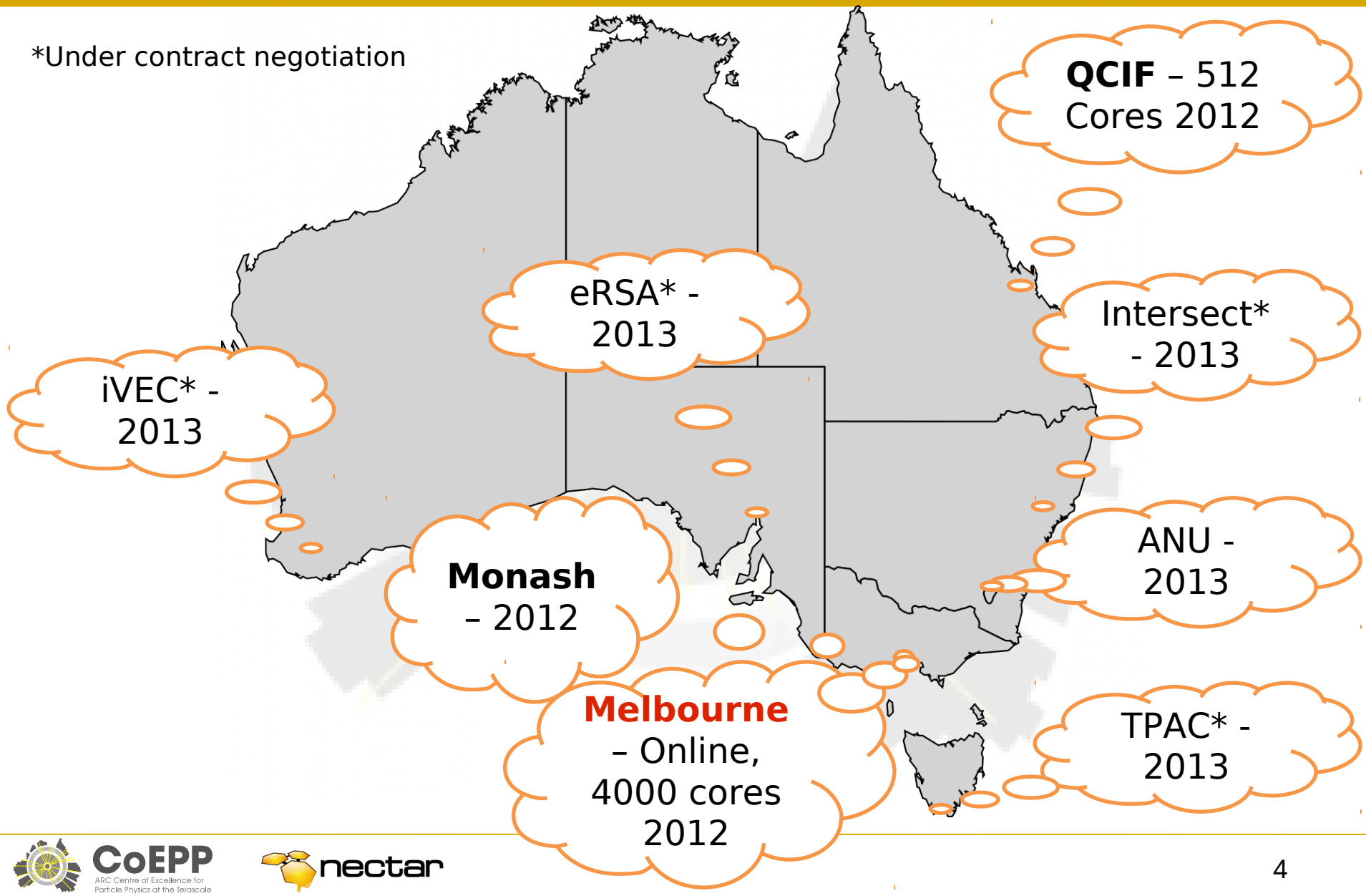
- NeCTAR Research Cloud
- Tier 2 - Integration with NeCTAR Cloud
- Tier 3 - Dedicated ATLAS Cloud-based Tier 3

NeCTAR and Its 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 25,000 cores Infrastructure-as-a-service cloud spanning 8 locations
  - Based on OpenStack

From One Node … to Eight in 2013

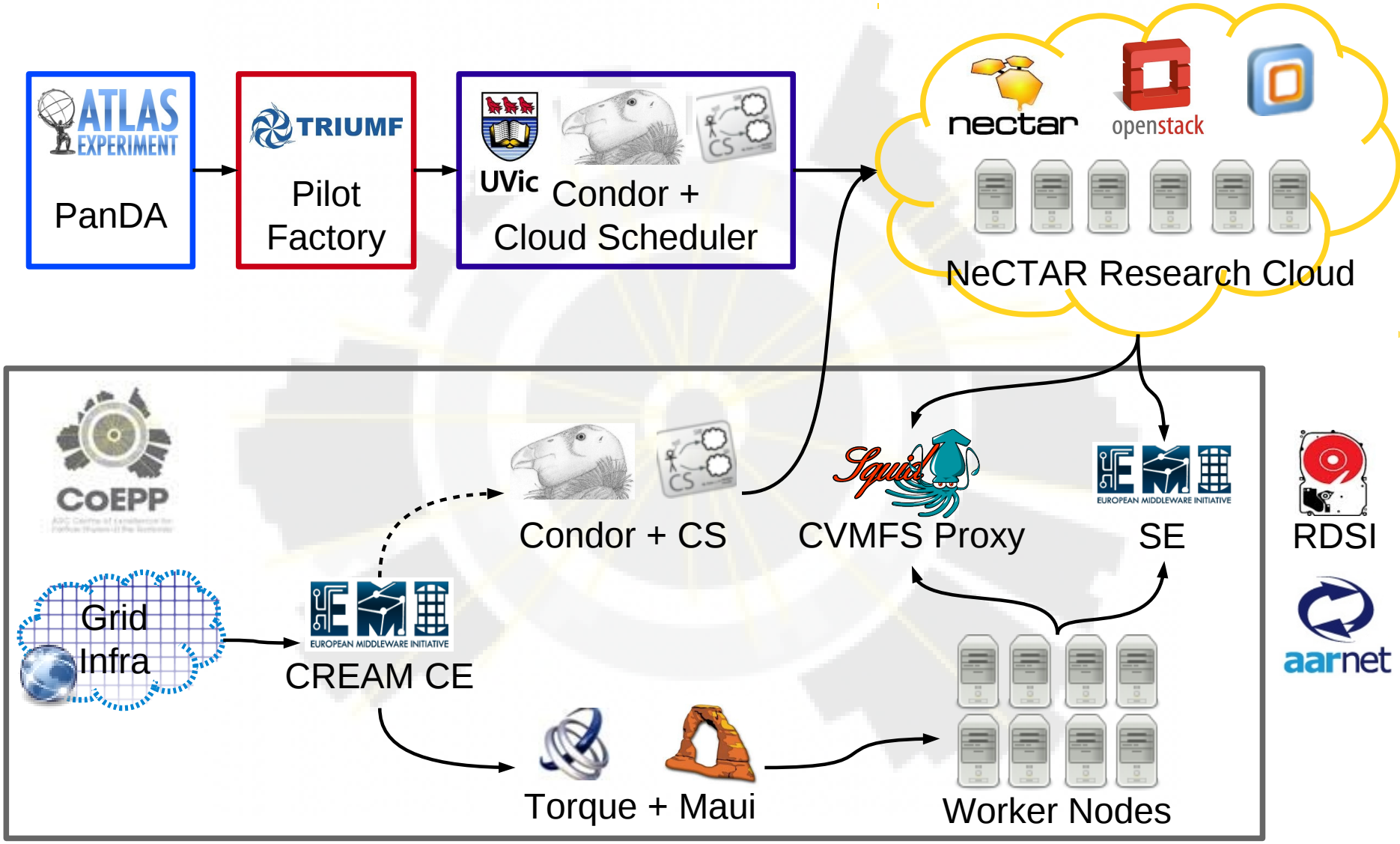
*Under contract negotiation



CoEPP Research Computing Cloud Project

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

Tier 2 System Framework



Tier 2 Requirements

- Phase I
  - Integrate with ATLAS PanDA job framework
  - Enable interoperability with OpenStack cloud
  - Dynamic scalability using Condor batch system and Cloud Scheduler software
  - CernVM Batch Node KVM image used
  - Subsequently run ATLAS MC production and user analysis jobs
  - Leverage the efforts from grid community

Tier 2 Requirements (cont.)

- Phase II
  - Extend computing capacity to 2000 cores
  - Extend storage capacity to 2PB by integrating Research Distributed Storage Infrastructures wherever possible (RDSI funded)
  - Support multiple VOs, e.g. Belle
  - Integrate with physical Tier 2 grid services, e.g. CREAM, DPM

Where Are We Now?

- A custom CernVM image for OpenStack
  - Recipe: https://rc.coepp.org.au/cernvm_nectar
- PanDA production and analysis queues:
  - Australia-NECTAR and ANALY_NECTAR
- 160 cores running ATLAS production jobs
  - 20 8-cores machines on NeCTAR cloud

Tier 2 Challenges

- Problems booting from CernVM image with OpenStack Essex
  - Error hasn't been seen since upgrade to Folsom
- Reliability issues with the underlying filesystem on NeCTAR cloud, e.g. inaccessible VMs, slowness
  - An undiagnosed issue, investigation on going
- The on-instance ephemeral storage is presented as a raw block device with no partition table or filesystem
  - Created an init script in CernVM image to partition, format and mount disk before Condor is up

Tier 2 Challenges (cont.)

- No DNS provisioning for VMs on NeCTAR, thus OpenStack reports wrong public hostname and IP address to its interfaces
  - Fixed Condor init script on CernVM image and patched Cloud Scheduler service
- CREAM Condor module is out-of-support, i.e. no updates for security, BDII, APEL, etc.
  - Either maintain it ourself or look for an EMI-supported batch system which is suitable for dynamic nature of on-demand scalability of cloud

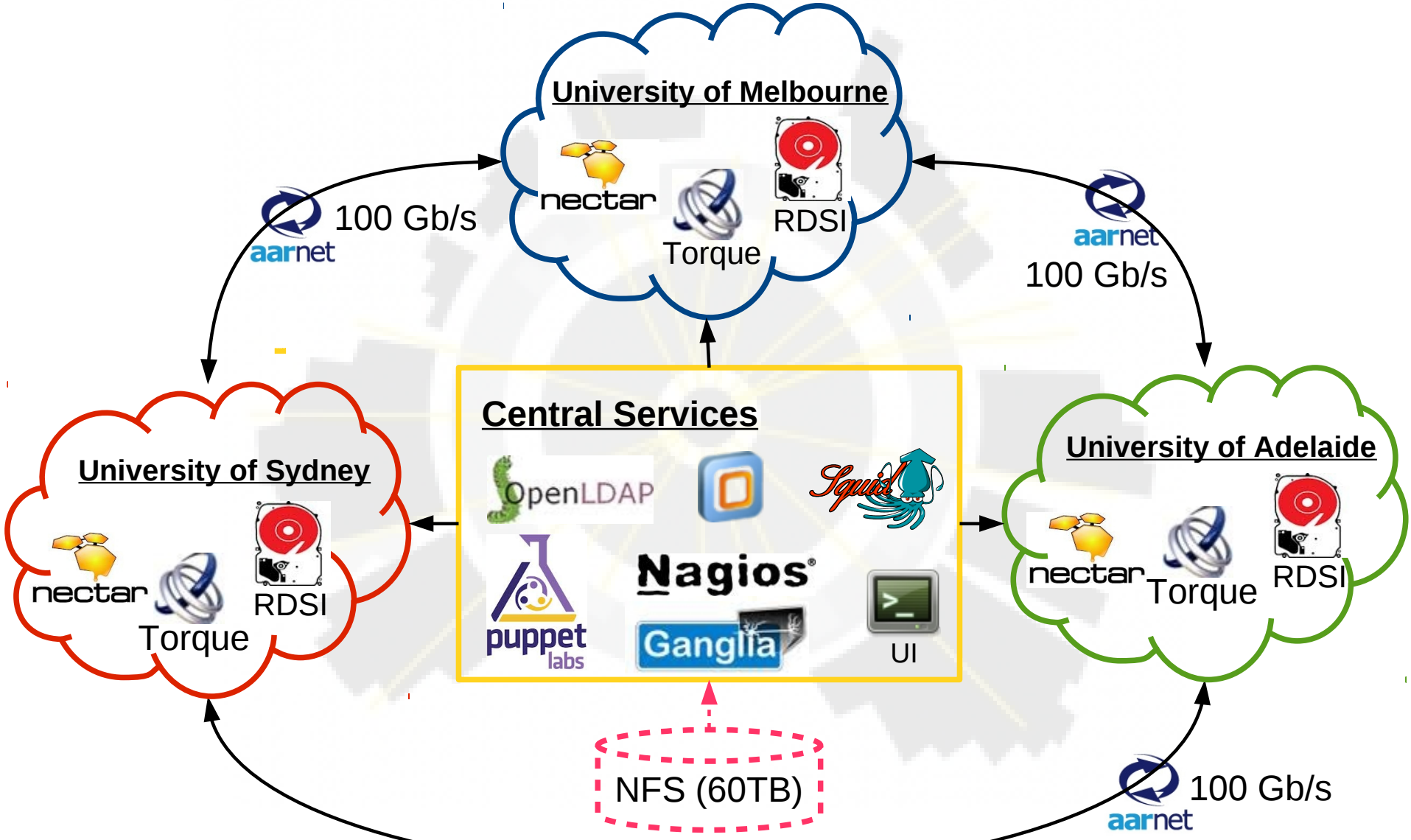
Tier 2 Challenges (cont.)

- Data locality issue with distributed storage
  - Collaborate with RDSI to work out the data presenting interfaces, desired filesystem architecture, etc.
  - Leverage 100Gb/s network connectivity provided by AARNet (Australian Academic and Research Network)

Tier 2 Future Work Plans

- Improve filesystem performance on batch node
- Deploy our own Condor batch server and Cloud Scheduler service in Australian Tier 2
- Integrate with current Puppet automated deployment and configuration management system in Tier 2
- Investigate a solution to integrate with CREAM
- Define technical requirements for external RDSI storage integration

Tier 3 System Framework



Tier 3 Requirements

- Federated and cloud based Tier 3 made up of resources at each of Adelaide, Melbourne, and Sydney
- Central services for software repository, authentication and authorisation, automated deployment, configuration management, monitoring
- Distributed computing and storage resources by integrating with NeCTAR and RDSI funded infrastructures

Where Are We Now?

- A base VM image: Scientific Linux 5 + Puppet
- Shared central services
  - **Puppet Master** as a central automated deployment and configuration management system
  - **UI** for users to submit batch jobs
  - **TORQUE** + **Maui** to control batch queues, resources and job scheduling
  - **NFS Server** to centrally locate users home directories
  - **Kerberos** and **LDAP** for federated user authentication and authorisation
  - **CVMFS Server** and **CVMFS Proxy** to host and distribute softwares
  - **Nagios** and **Ganglia** for service availability and performance monitoring
- Nodes on the cloud
  - 120 single-core TORQUE batch nodes
  - 2 16-cores interactive nodes

Where Are We Now? (cont.)

- In-house tools to streamline operations and automation
- Ability to build entire infrastructure on cloud within minutes by just typing one command

Tier 3 Challenges

- The first setup on OpenStack Essex environment was initially unstable.
  - Created tools to check system health and automated fix problems as found
- Experienced many delays due to OpenStack Essex instabilities and bugs e.g. Filesystem I/O errors, read-only filesystem, booting errors, network unreachable errors
  - Due to time constraints these problems were usually fixed with a terminate and reboot
- Problems reading custom SL5 KVM image with OpenStack Essex
  - Issue hasn't been seen since upgrade to Folsom

Tier 3 Challenges (cont.)

- Allocated resources reserved error with OpenStack Folsom
  - Asked NeCTAR to free up resources for now
- On-demand scalability of Tier 3 cloud
  - Standard TORQUE is not suitable for dynamic batch node provisioning
  - Investigating approaches for dynamically scale-out and scale-in batch instances across distributed computing resources on cloud
- Single job submission entry point with resource discovery ability
- Integration with distributed filesystem

Tier 3 Future Work Plans

- Separate Puppet production and testing environment
- Investigate resource discovery solutions to provide truly federated system with single job submission entry mechanism
- Search for an integration solution which empower TORQUE cluster to manage dynamic computing resources on cloud environment
- Build a pilot system with RDSI storage backend and DPM filesystem headhode

Backup Slides

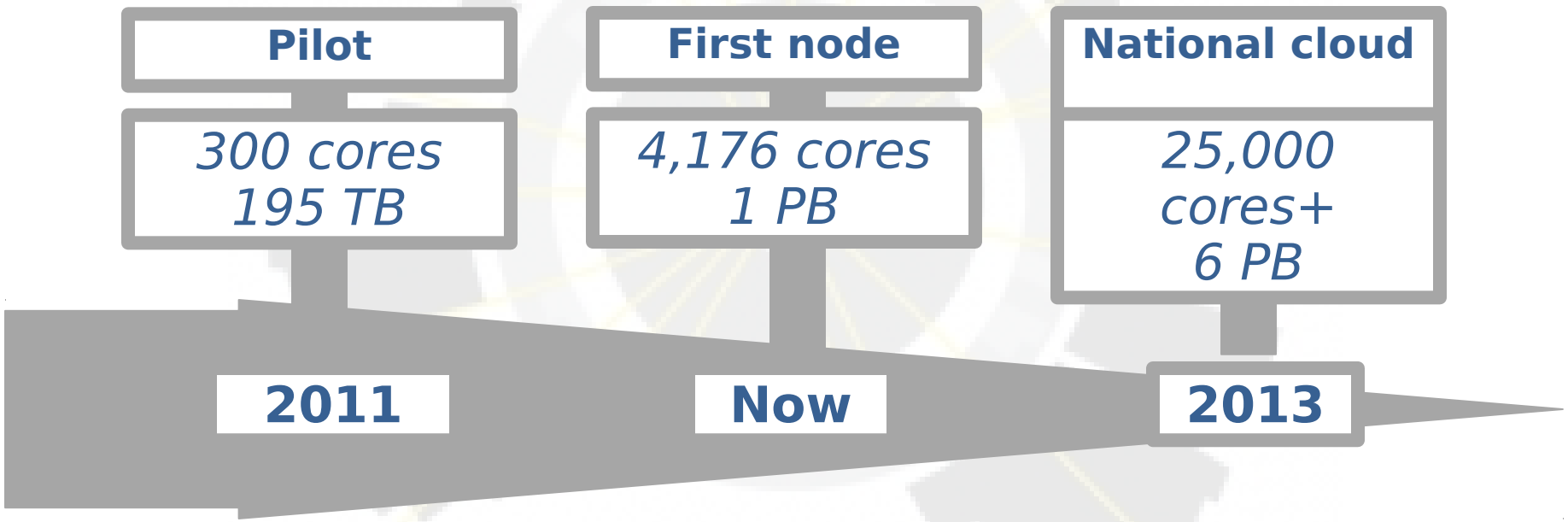
Specifications – University of Melbourne Cloud Node

- OpenStack Folsom/Stable (+~5% Grizzly backport)
  - Ubuntu 12.04 LTS, KVM, Puppet
- Hardware
  - 336 cores – 48 Core Dell R815s
  - 3840 cores – 160* 24 core, 128GB, 10Gbit/s Xenon Quad2U
  - 195TB – HP DL180G w/ DL2000 @ 24TB/node
  - 146TB – Dell R715 w/ MD1200 @ 24TB/node
  - 10Gbit/s – Cisco Nexus (2232, 5596, …)
  - Hitachi HNAS/BluARC – 100TB for running VMs

NeCTAR Research Cloud... Why Build It Ourselves?

- Proximity – the honeypot –infrastructure attracts community
- Local infrastructure is more responsive to research needs
- Service offering and usage modes suitable for research
- Locality to instruments, research networks and other infrastructure.
- Data sovereignty

NeCTAR Research Cloud Timeline



Overall Timeline - Infrastructure Extension