SCA19 APRP

Update

Andrew Howard - Co-Chair APAN APRP Working Group

NCI AUSTRALIA Overview

- What is a Research Platform
  - Notable Research Platforms
- APRP
  - History
  - Participants
  - Activities

  - We live in an age of rapidly expanding data growth

Types of data transfer

Real Time
Best effort
Streaming

90% of the worlds data was created in the past 2 years

What is a Research Platform

- Connects peak HPC, Cloud and Research facilities connected via 100G networks
- Focus on Data intensive science
- Supporting research and data movement for
  - Particle physics
  - Astronomy
  - Biomedical sciences
  - Earth sciences
  - Scalable data visualisation

Notable Research Platforms

- Pacific Research Platform
  - Built on the optical backbone of Pacific Wave, a joint project of CENIC and the Pacific Northwest GigaPOP (PNWGP) to create a seamless research platform that encourages collaboration on a broad range of data-intensive fields and projects.
  - 50+ institutions, led by researchers at UC San Diego and UC Berkeley.
  - Includes the National Science Foundation, Department of Energy, and multiple research universiti[illegible] the world.
- US National Research Platform

APRP participants

- KISTI - Korea
- NCI - Australia, CSIRO - Australia
- Perdana U - Malaysia Putra U - Malaysia
- NSCC - Singapore
- Tsinghua U & NSCC Wuxi - PRC
- Starlight USA

NRENs and Networks

- AARNet
- SingAREN
- KREONET
- Internet2
- Pacific Wave
- Starlight
- Transpac

APRP History

- BoF at APAN 45 and SCA18
- Working group ratified at APAN 46
- APAN 47
- Charter and Technologies
  - Friction Free Data Movement
  - European eXtreme Data Cloud
  - Problem space
  - Potential Solutions
  - Research Platforms
  - APRP proposed design

Australian National Research Platform

- Australia National Research Platform
  - NCI
  - Pawsey
- How it relates to APRP
- Foundation capabilities
  - Data Movement
  - Federated authentication
  - Service orchestration

Data Mover Challenge

- Data Mover Challenge
  - APRP participants
    - Australia
    - Singapore
    - Japan
    - Korea
    - USA

Foundation Capabilities - Data movement

- Data movement
  - File replication
  - Object replication
  - Scheduled and background transfers
- Service endpoints
  - Shared capabilities
  - Distributed data stores integrated into a single metadata namespace
- Build on advanced network capabilities

Foundation Capabilities - Containerised Toolkits

- Containers
  - Encapsulate common data transfer workflows
    - Globus/gridftp
    - http
    - Big Data Express
    - Other DMC toolkits
  - Bio-Informatics
    - Galaxy

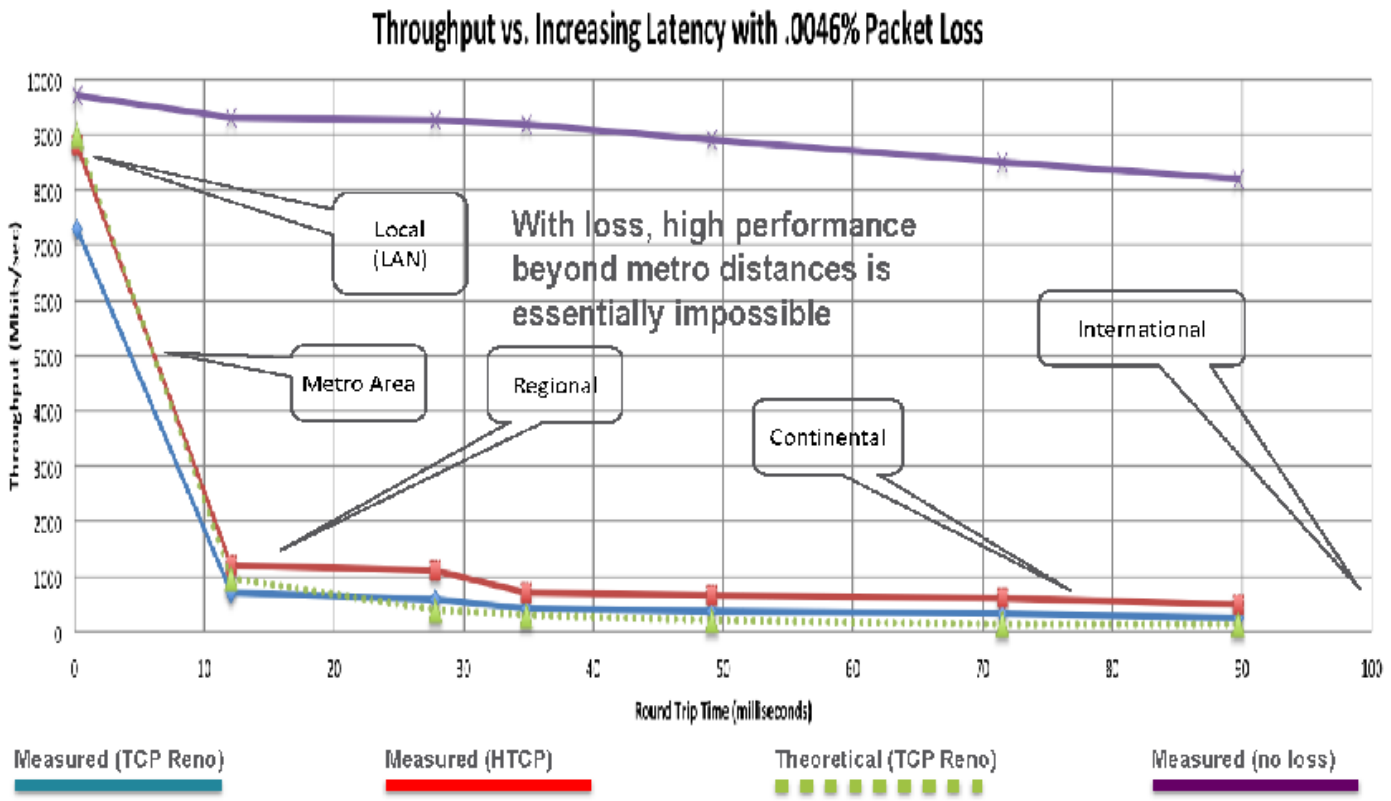
Friction Free Data movement

- We need to provide our researchers with a friction free data transfer system
  - Easy to use
  - Secure using a Federated Access system
- The network and tools should have the data in the right location at the right time
- Able to effectively use different storage tiers
  - SSD
  - Spinning Disk
  - Tape
- The researcher creates a Data Intent definition
  - Data Source
  - Data Target
    - Transfer priority (High, Medium, Low)
    - Storage performance (SSD, Disk, Tape)
  - optional Network intersection

TCP/IP

By default TCP/IP does not perform well over high bandwidth, high delay circuits.

A small amount of packet loss makes a huge difference in TCP performance



The challenges of the Big Data era

How to… ?

... orchestrate and federate Cloud, Grid and HPC [public or private] resources?

... Avoid software and vendor lock-in?

... overcome performance issues limiting massive adoption of virtualised Cloud resources in large data centres?

... exploit specialised hardware, such as GPUs or low-latency interconnections?

... manage dynamic and complex workflows for scientific data analysis?

… combine data from multiple sources and stored in multiple locations through incompatible technologies?

… support federated identities and provide privacy and distributed authorisation in open Cloud platforms?

... provide APIs to exploit the above and write applications, customisable portals and mobile views?

... move beyond statical location and partitioning of both storage and computing resources in data centres?

... distribute and deploy applications in a flexible way?

... exploit distributed computing and storage resources through transparent network interconnections?

Capabilities and Requirements

- Regional connection
- Federated access
- Data capacitor capabilities
  - Local storage
- Container provisioning
  - Instantiate toolkit containers
- VM provisioning
  - Provide VM access on regionally connected DTN

Containers

- Containers
  - Docker in a well protected hosting environment
  - Singularity
    - V2
    - V3
  - Lightweight services

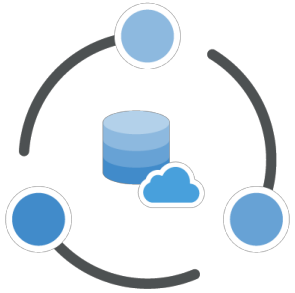
Role of NRENs

- Our National Research and Education Networks are critical
  - Advanced network services
    - 100G
    - Anycast
    - IPV6
  - Data sharing services (AARNet Cloudstor)
  - National service termination point

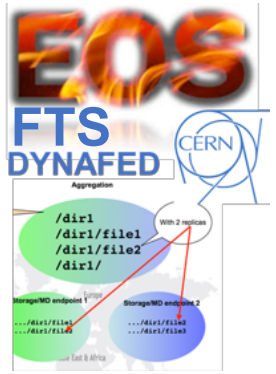
Open Science Data Cloud

eXtreme DataCloud

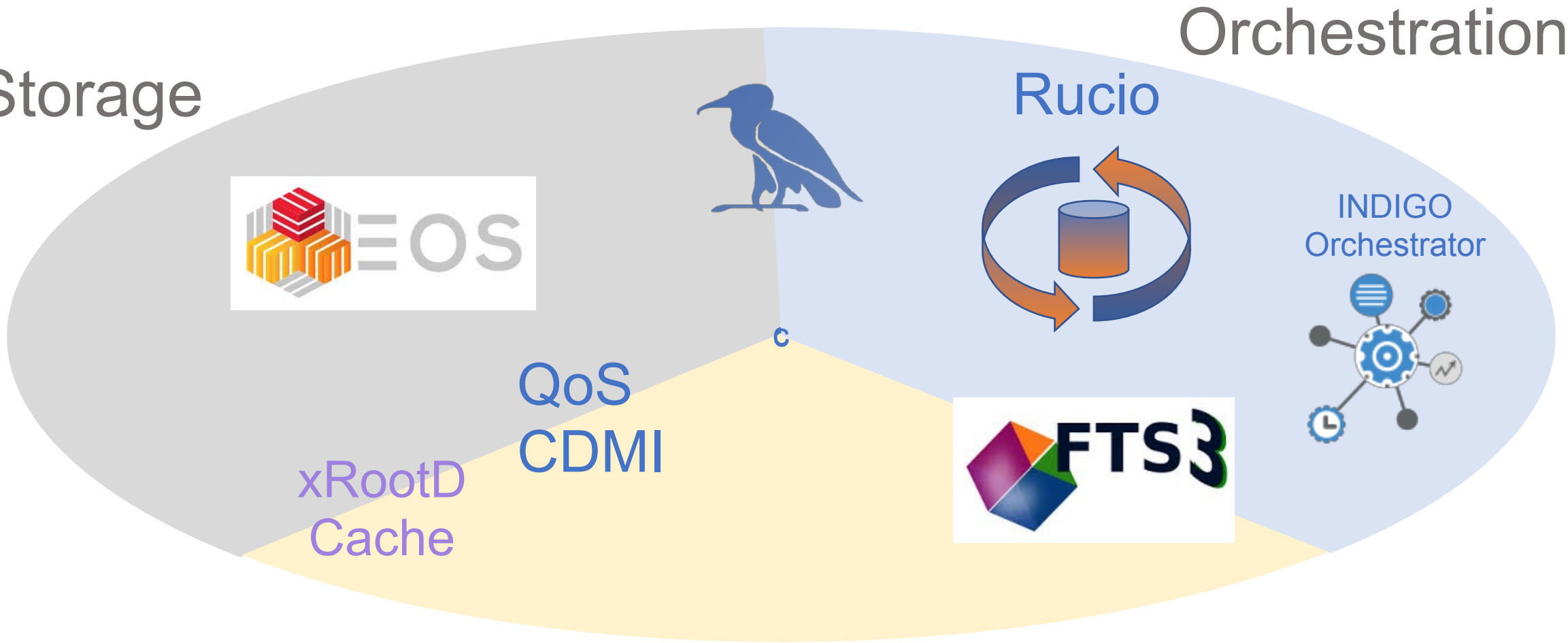
INDIGO PaaS Orchestrator



INDIGO CDMI Server



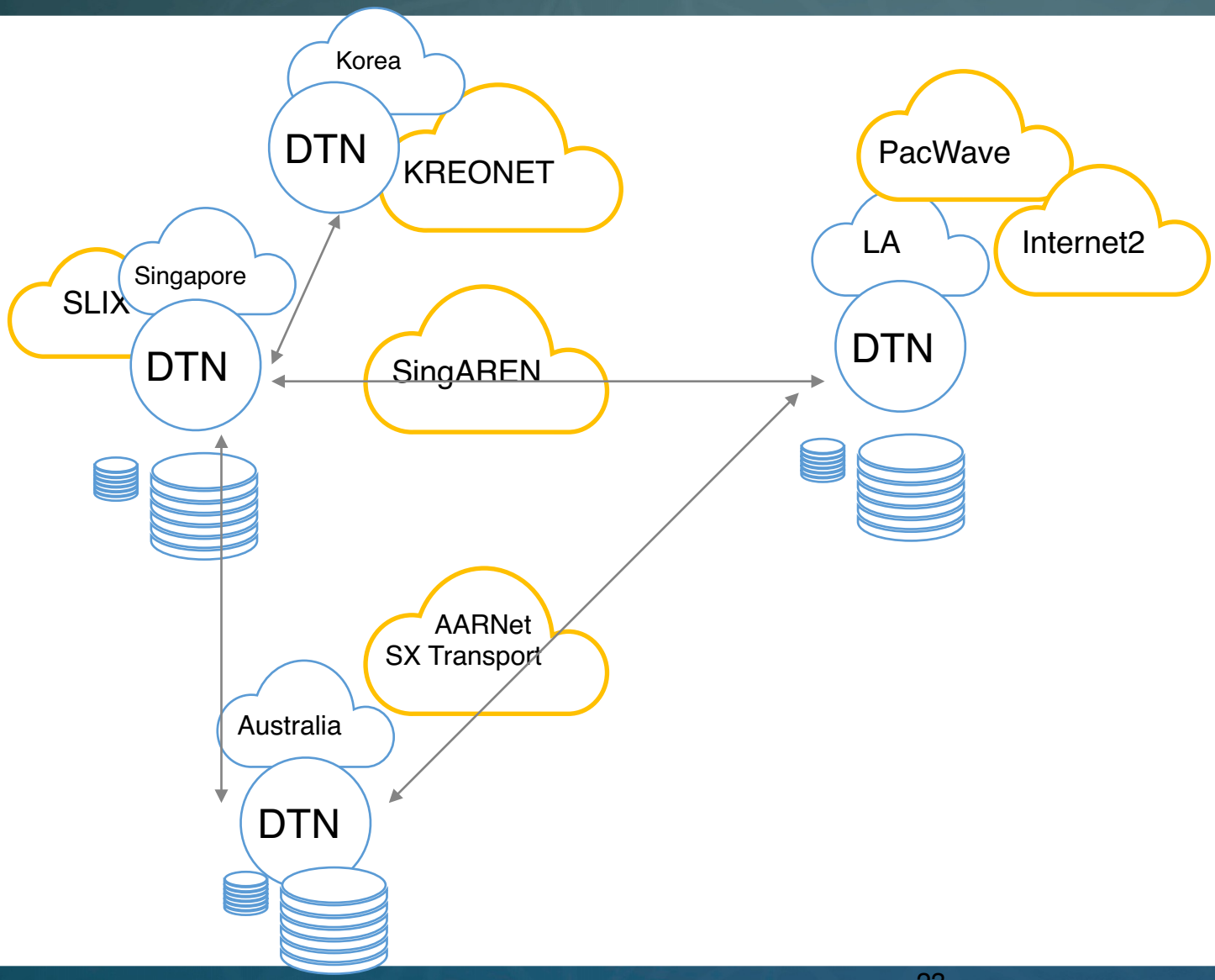
XDC components



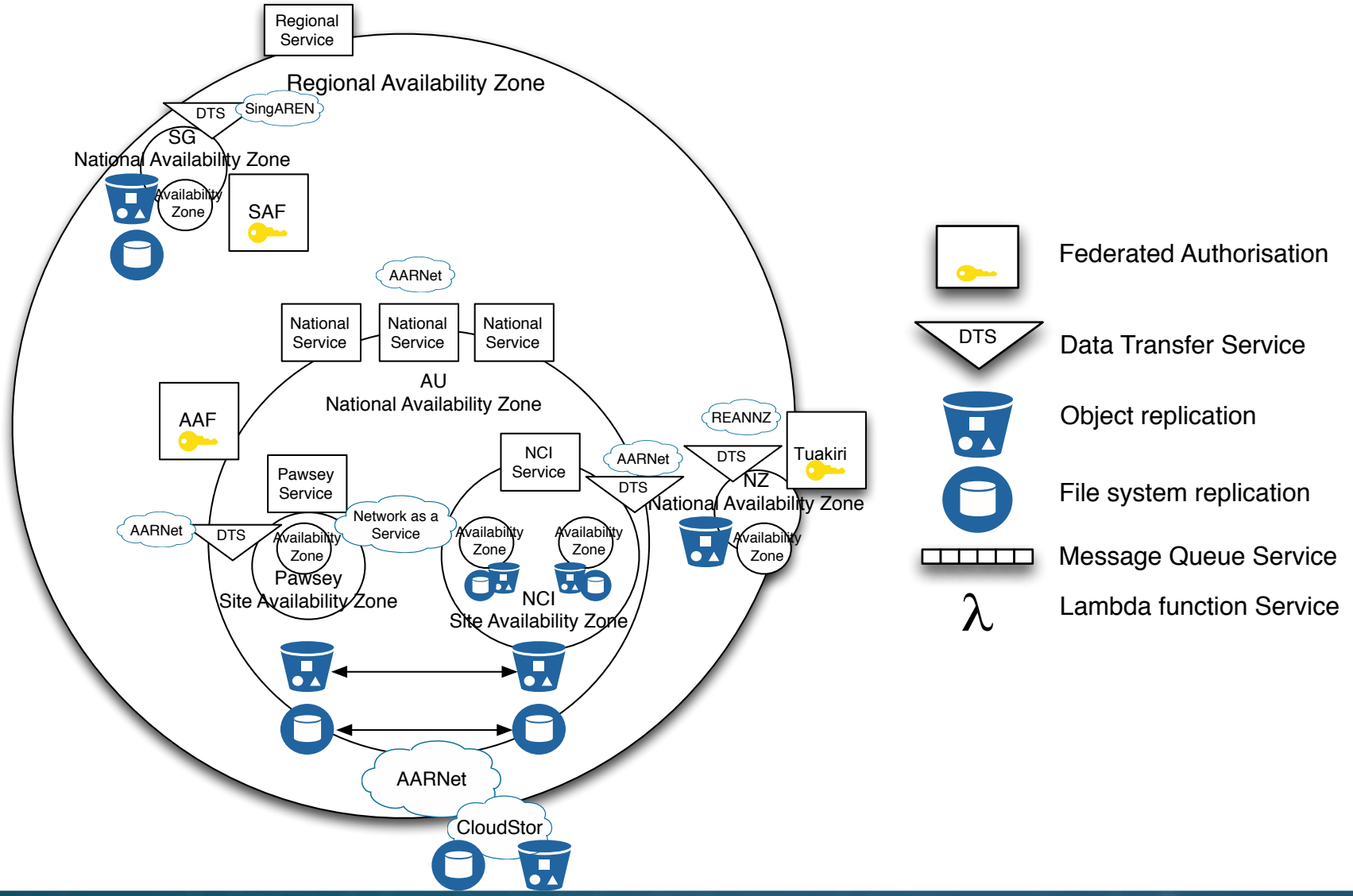
Research Platforms

- Pacific Research Platform (PRP)
  - US Initiative to build a network of Science DMZs with well tuned systems for data movement
- Asia Pacific Research Platform (APRP)
  - Regional initiative
    - KISTI - Korea, NCI - Australia, Perdana U - Malaysia, NSCC - Singapore, Tsinghua U & NSCC Wuxi - PRC, CSIRO - Australia, Putra U - Malaysia

APRP proposed high level design



National and Regional Research Platform Architecture



Capabilities and Requirements

- Regional connection
- Federated access
- Data capacitor capabilities
  - Local storage
- Container provisioning
  - Instantiate toolkit containers
- VM provisioning
  - Provide VM access on various National Research and Commercial clouds

High level architecture/goals

- Services may operate at a Site, National or Regional scope.
- Replication of Objects and Filesystems to support services operating in multiple Availability Zones.
- Authentication support for existing LDAP based systems and Federated identities through AAF and other federated Federations (eduGain).
- Share common best practice and personnel in design and implementation.
- Efficiently support the rapidly growing national BioInformatics activities.

Capability required

- GPU access
- Object store replication
- File system replication
- Data transfer services
- Advanced Cloud development testbed
- Containers
- Message queues
- Functions

Conclusion

- We have started the journey
- The foundation of data movement is in progress
- Activities like the DMC are building better collaboration and co-
- We need to investigate other shared resources
- We invite participation
- Global Research Platform

Contact details

- For more information please contact me

  andrew.howard@anu.edu.au

Questions ?


thank you
danke
рахмат
謝謝
spas
ngiyabonga
شكراً جزيلاً
Баярлалаа
mersi
kia ora
barka
welalin
tack
teşekkür ederim
mahalo
tapadh leat
спасибо
faafetai lava
vinaka
спасибі
blagodaram
dank je
misaotra
matondo
paldies
grazzi
gracias
хвала
asante
manana
kiitos
dankie
nanni
nandri
dhanyavad
mauruuru
köszönöm
aciu
dankon
akun
obrigada
tenki
murakoze
enkosi
bedankt
bayarlalaa
gràcie
hvala
djiere dieuf
tau
mochchakkeram
дякую
mamnun
dziękuję
chnorakaloutioun
gratias ago
gràcies
sulpáy
taiku
go raibh maith agat
chokrane
sobodi
děkuji
sağolun
sukriya
kop khun krap
arigatô
takk
dakujem
trugarez
obrigado
mèsi
didi madloba
najis tuke
ありがとう
tanemirt
rahmet
grazie
diolch
dhanyavadagalu
shukriya
mercé
мерси
kam sah hamnida
rahmat
terima kasih
감사합니다
xièxie
ευχαριστώ
merci

Acknowledgements