



How do we architect networks to accommodate “Big Science” projects of the future

Charles Smith - chas@cisco.com

**Cisco Academic & Research Technology
Initiatives Group - ARTI**

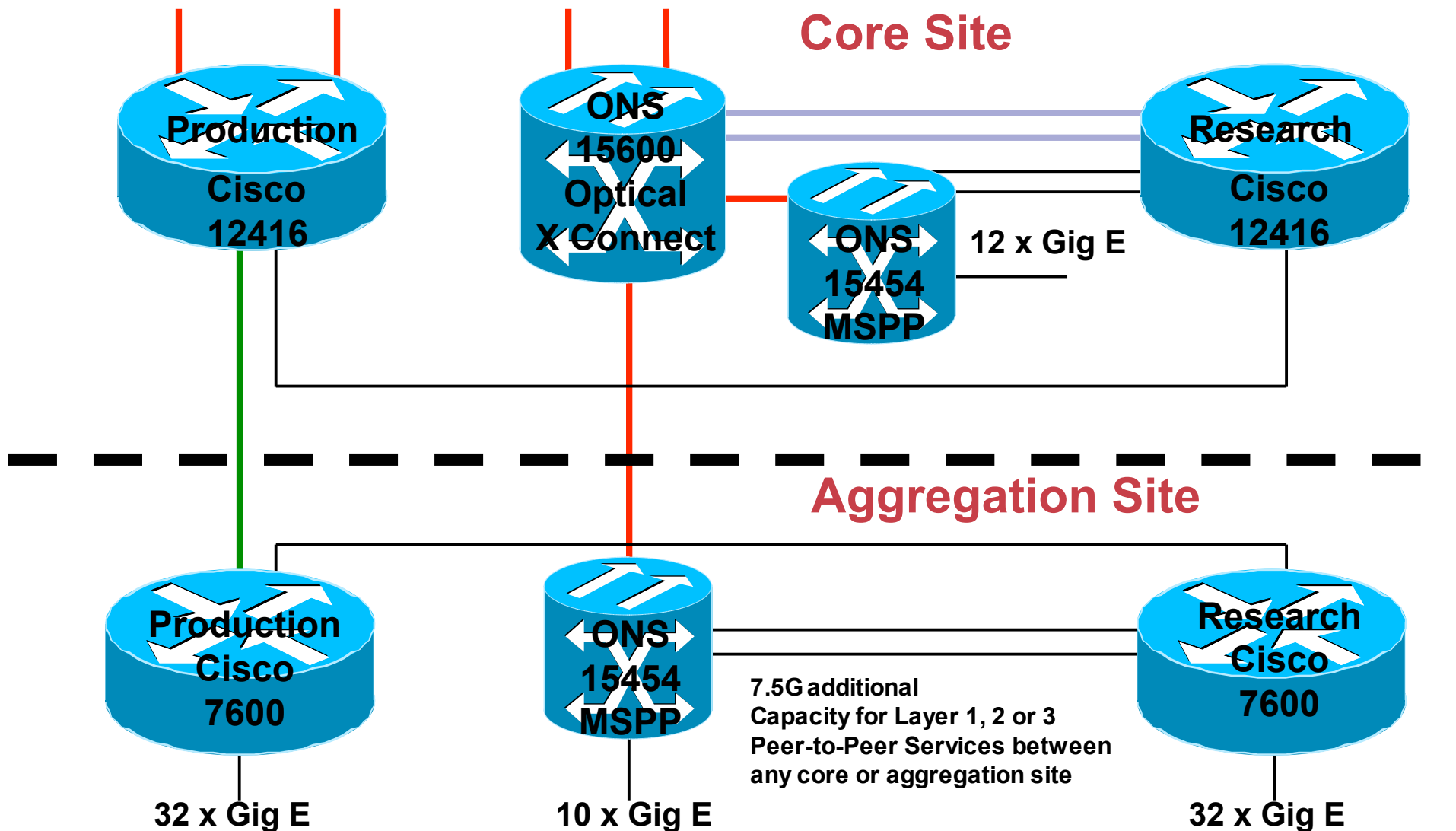
Today's NREN Backbones (Some Examples)

TWAREN

- **Production STM-64 Backbone**
 - L3 IPv6/IPv4 Backbone**
- **Research STM-64 Backbone**
 - L3 STM-16 MPLS/IPv6/IPv4 Backbone**
 - L2 Gigabit Ethernet Backbone**
 - L1 Gigabit Ethernet Lambda Backbone**

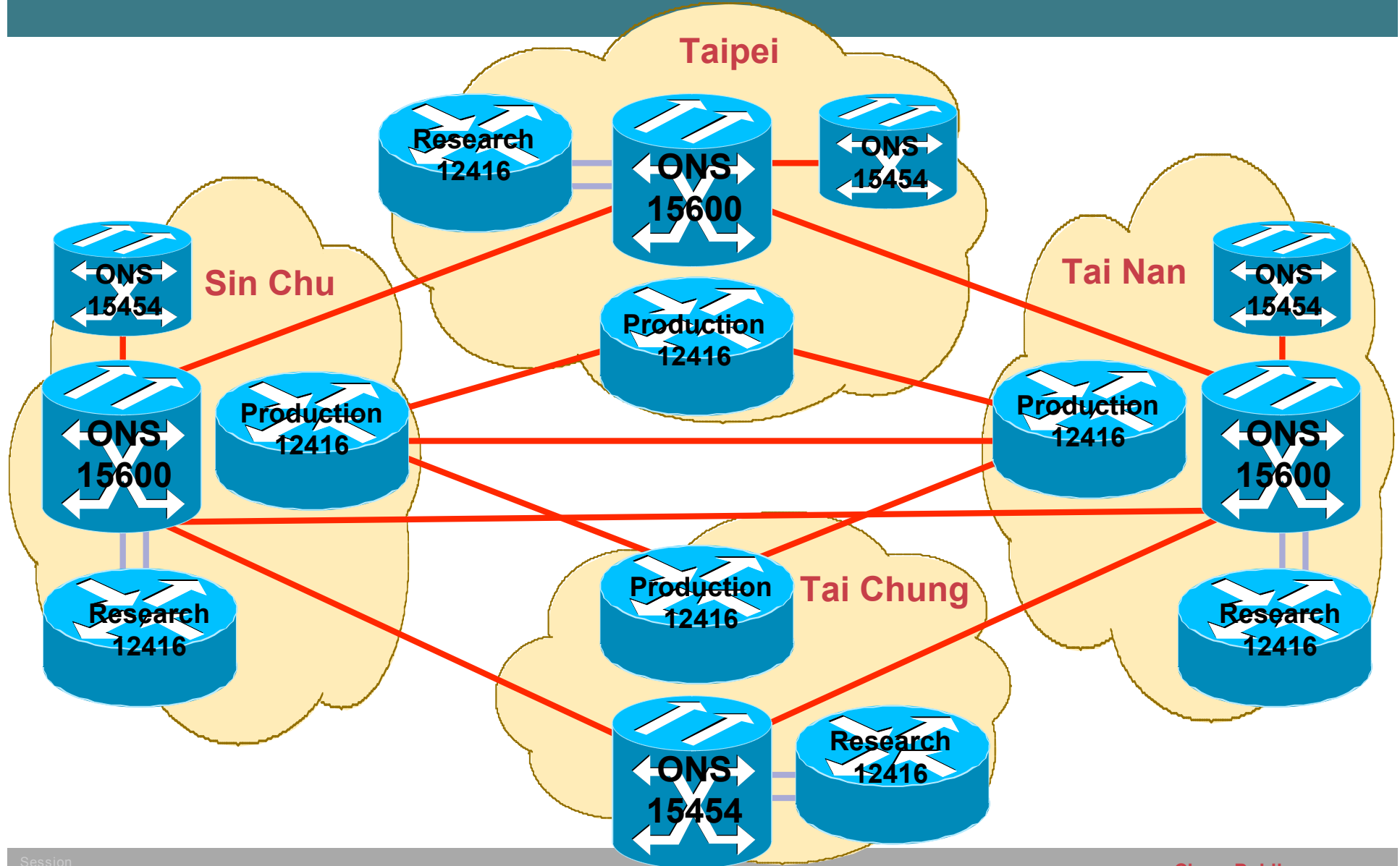
TWAREN Site Architecture

- STM-64/OC192
- 10 Gigabit Ethernet
- STM-16/OC48
- Gigabit Ethernet



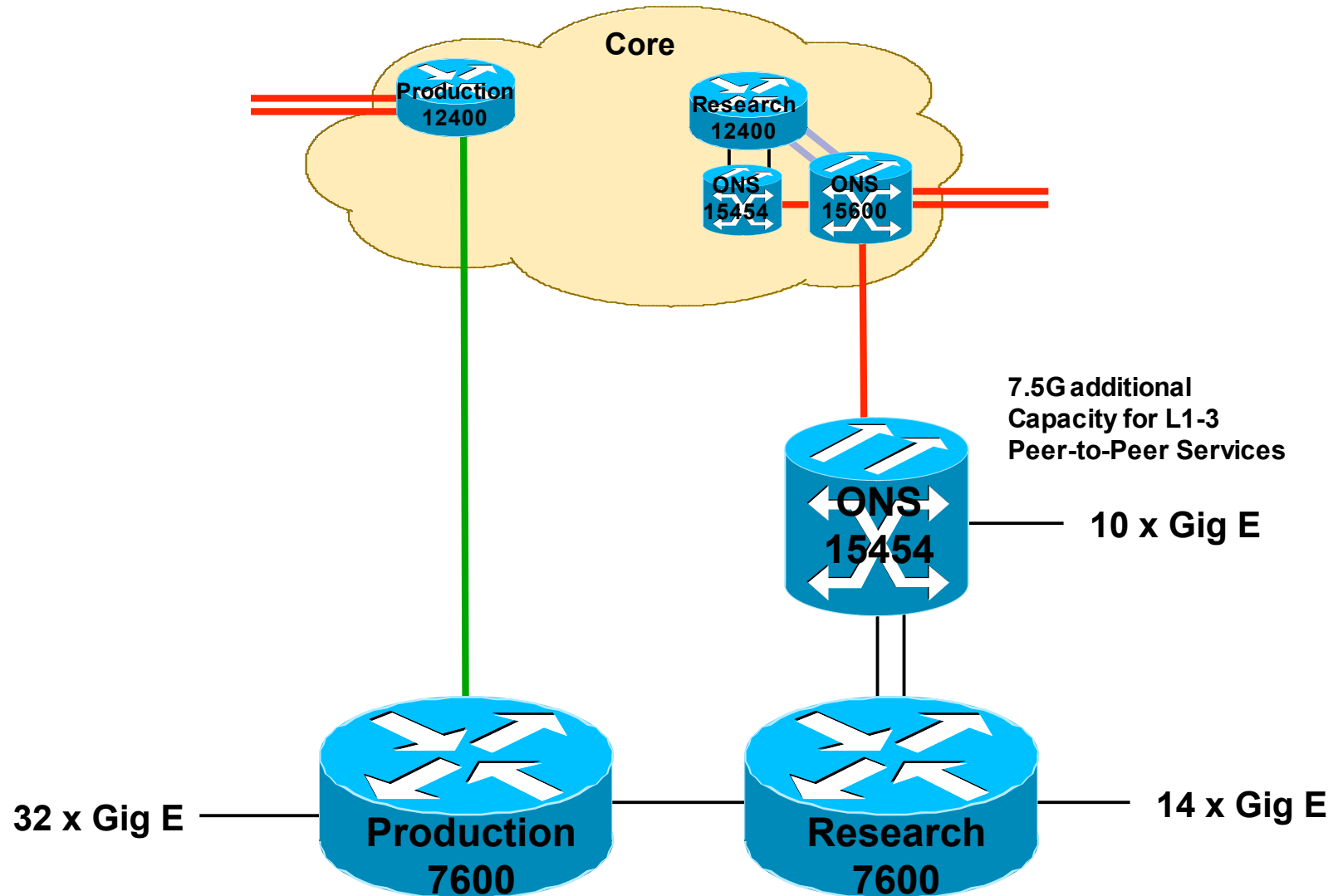
TWAREN Core Sites

- STM-64/OC192
- 10 Gigabit Ethernet
- STM-16/OC48
- Gigabit Ethernet



TWAREN Aggregation Site

- STM-64/OC192
- 10 Gigabit Ethernet
- STM-16/OC48
- Gigabit Ethernet



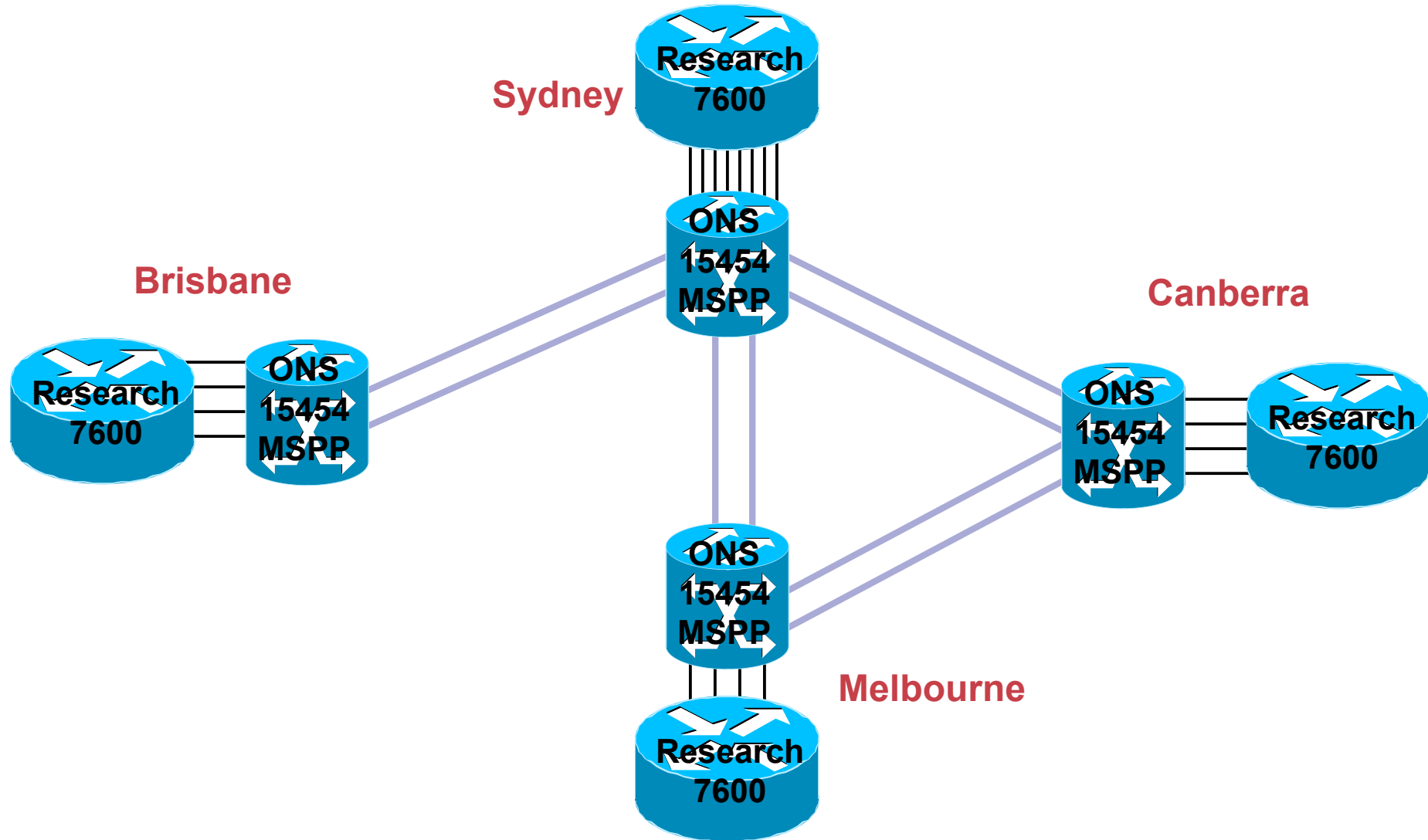
Today's NREN Backbones (Some Examples)

GrangeNET

- **Production 2xSTM-16 Backbone**
 - L3 IPv6/IPv4 Gigabit Ether Channel Backbone**
 - L2 Gigabit Ethernet Backbone**
 - L1 Gigabit Ethernet Lambda Backbone**

GrangeNet Site Architecture

— STM-16 Circuit
— Gigabit Ethernet



Today's NREN Backbones (Some Examples)

CENIC

- **Production DWDM Backbone**

 - L3 IPv6/IPv4 10 Gigabit Ethernet Research Backbone**

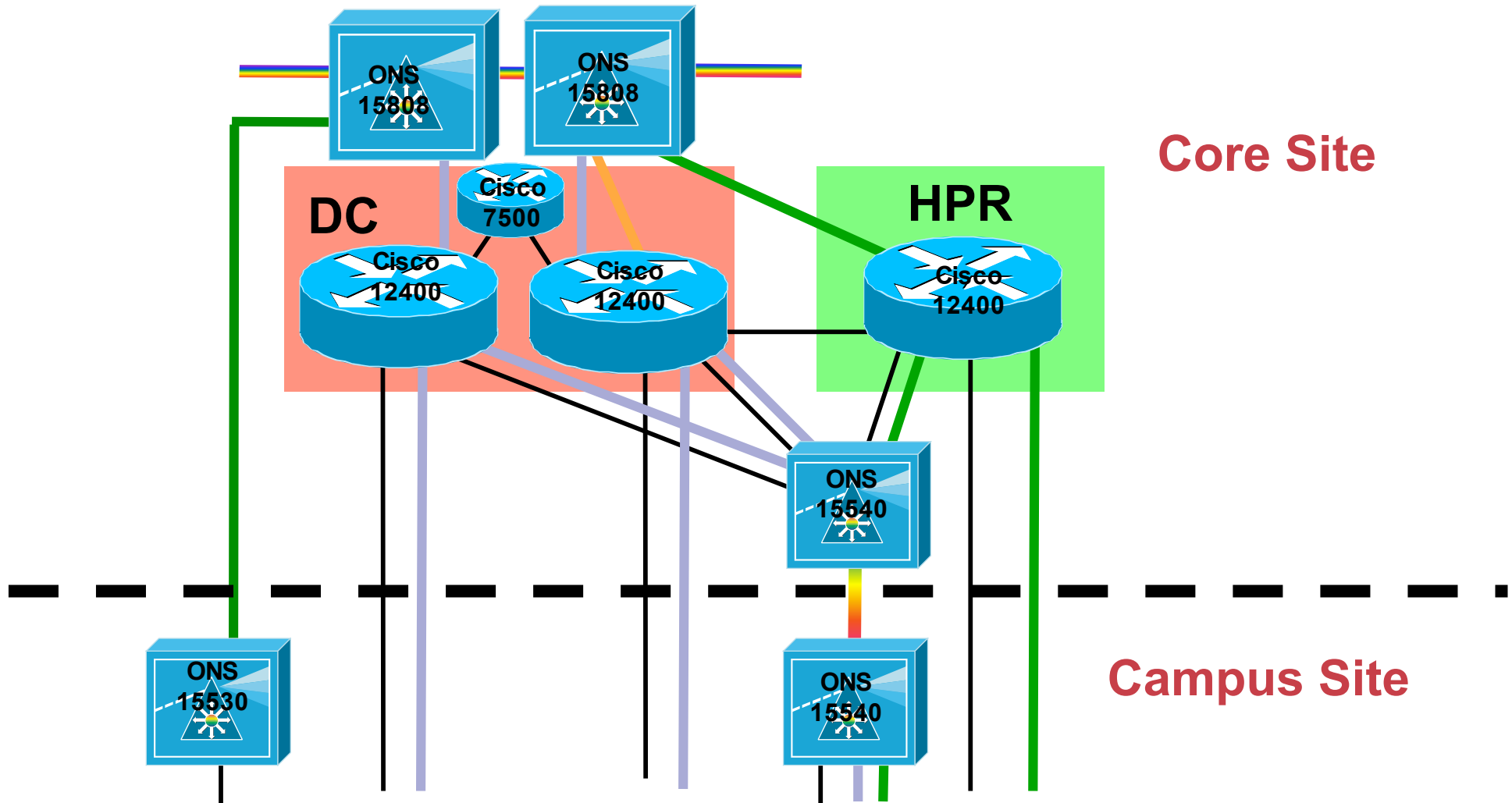
 - L3 IPv4 OC48 Production Backbone**

 - L2 10 Gigabit Ethernet Production Backbone**

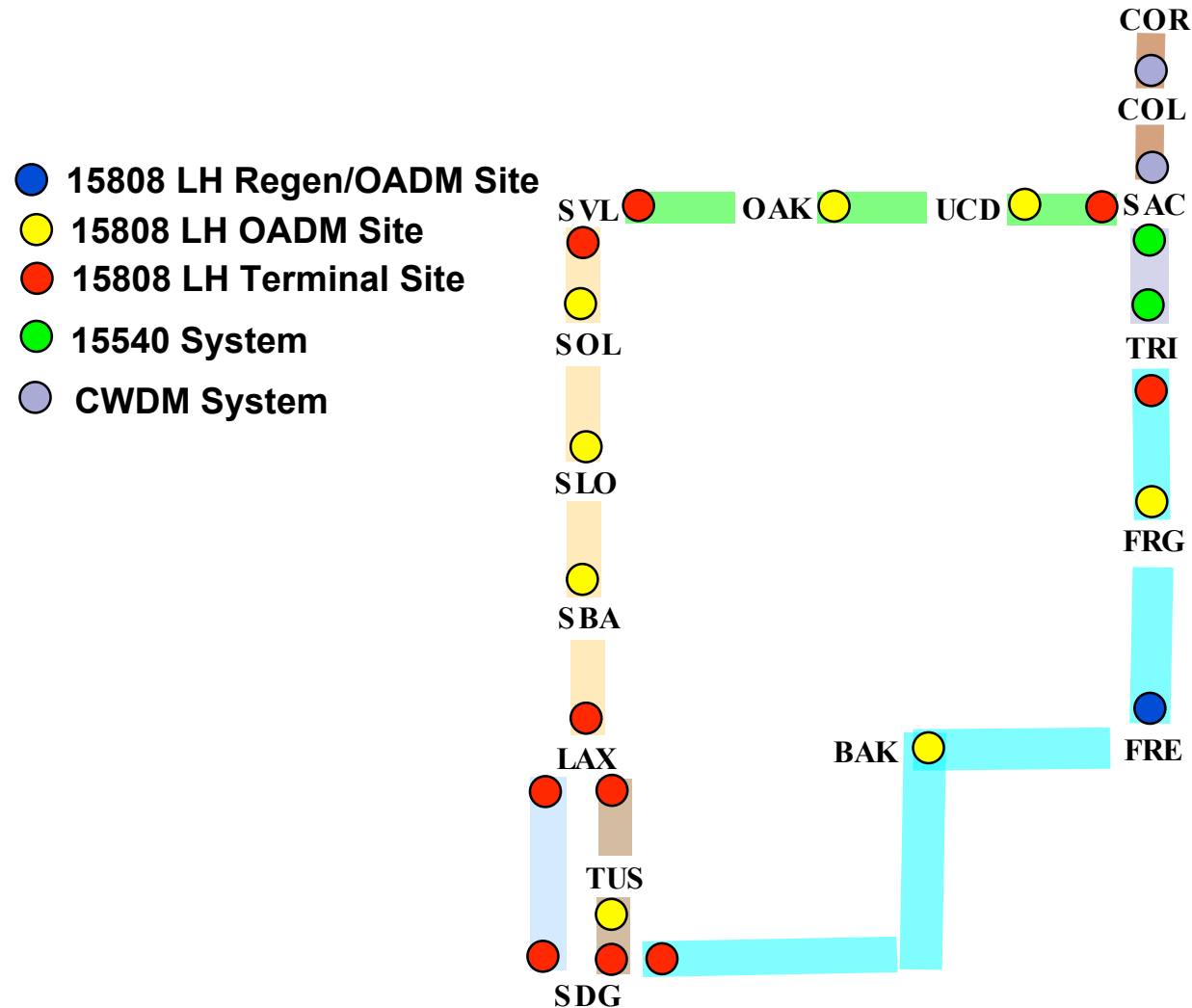
 - L1 10 Gigabit Ethernet Lambda Backbone**

CENIC Site Architecture

- STM-64/OC192
- 10 Gigabit Ethernet
- STM-16/OC48
- Gigabit Ethernet
- DWDM Lambdas

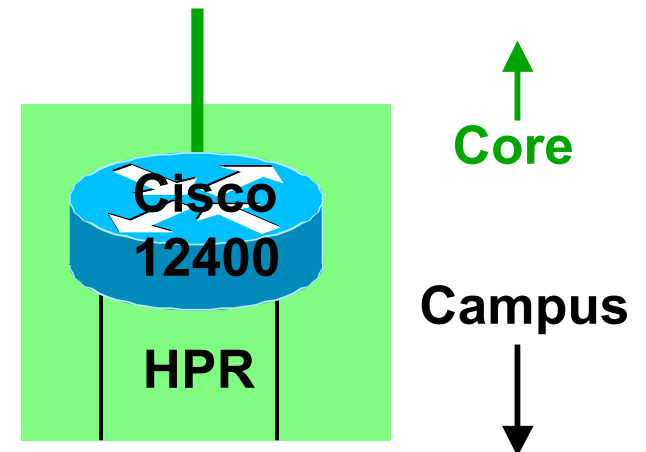
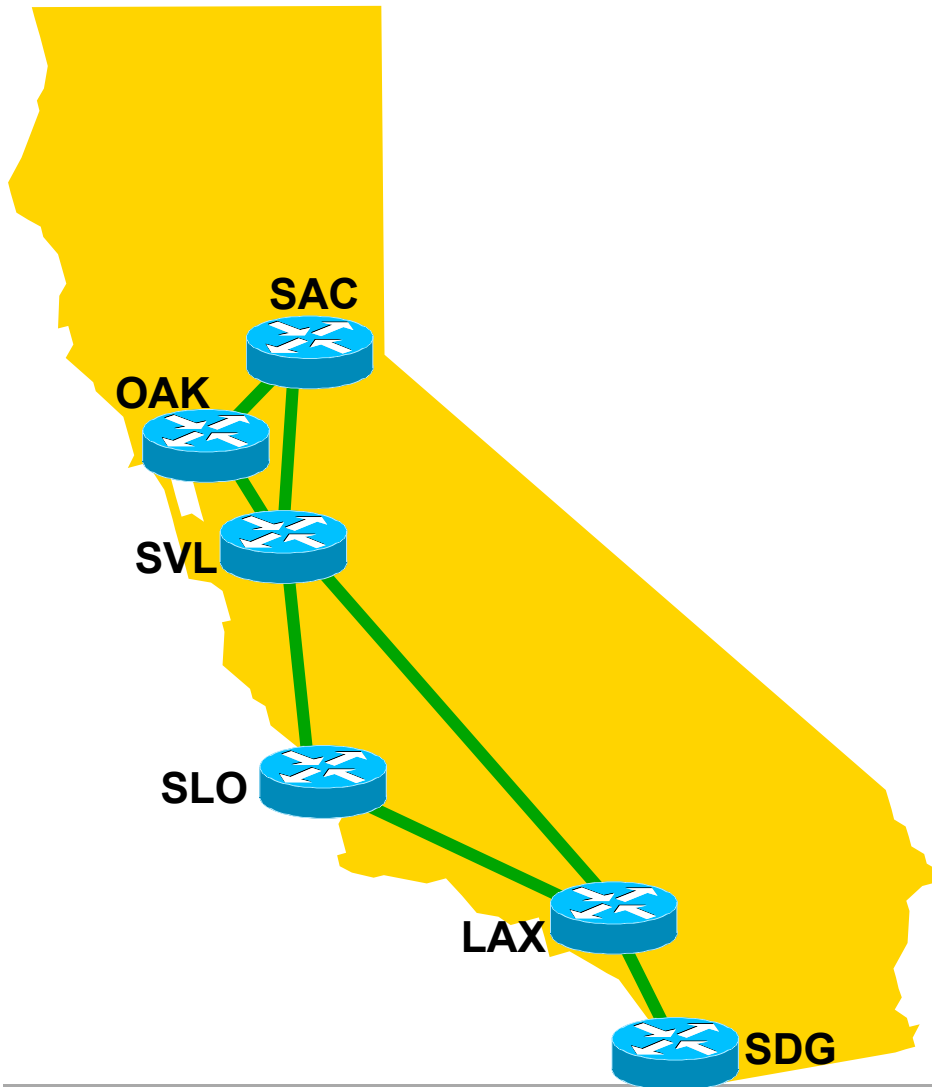


CENIC DWDM Backbone



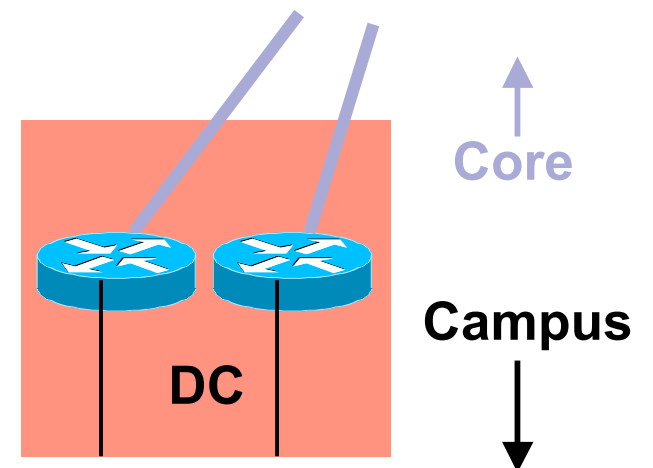
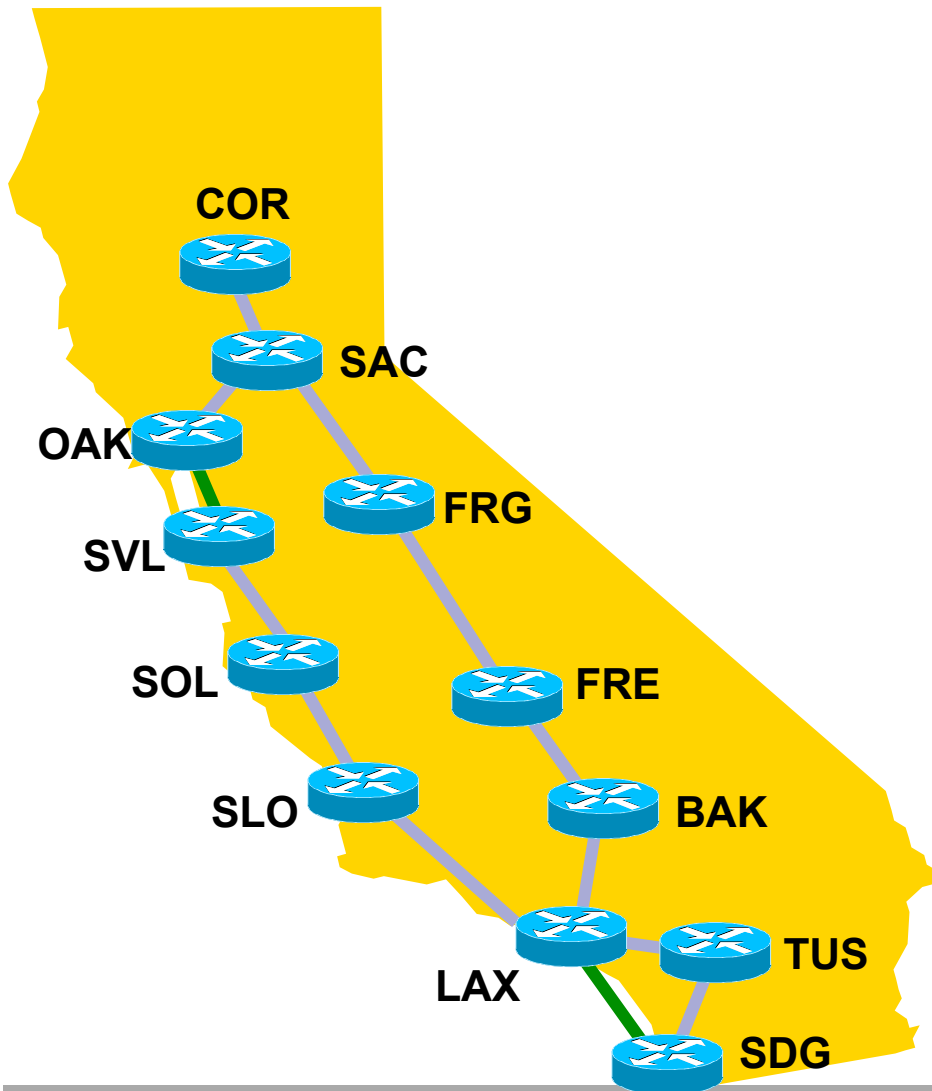
CENIC HPR Research 10GbE L3 IPv6/IPv4 Backbone

- STM-64/OC192
- 10 Gigabit Ethernet
- STM-16/OC48
- Gigabit Ethernet



CENIC Calren DC Production OC48 L3 IPv4 Backbone

- STM-64/OC192
- 10 Gigabit Ethernet
- STM-16/OC48
- Gigabit Ethernet



Today's NREN Backbones (Some Examples)

National Lambda Rail - NLR

- **Production DWDM Backbone**

 - L3 IPv6/IPv4 10 Gigabit Ethernet Research Backbone**

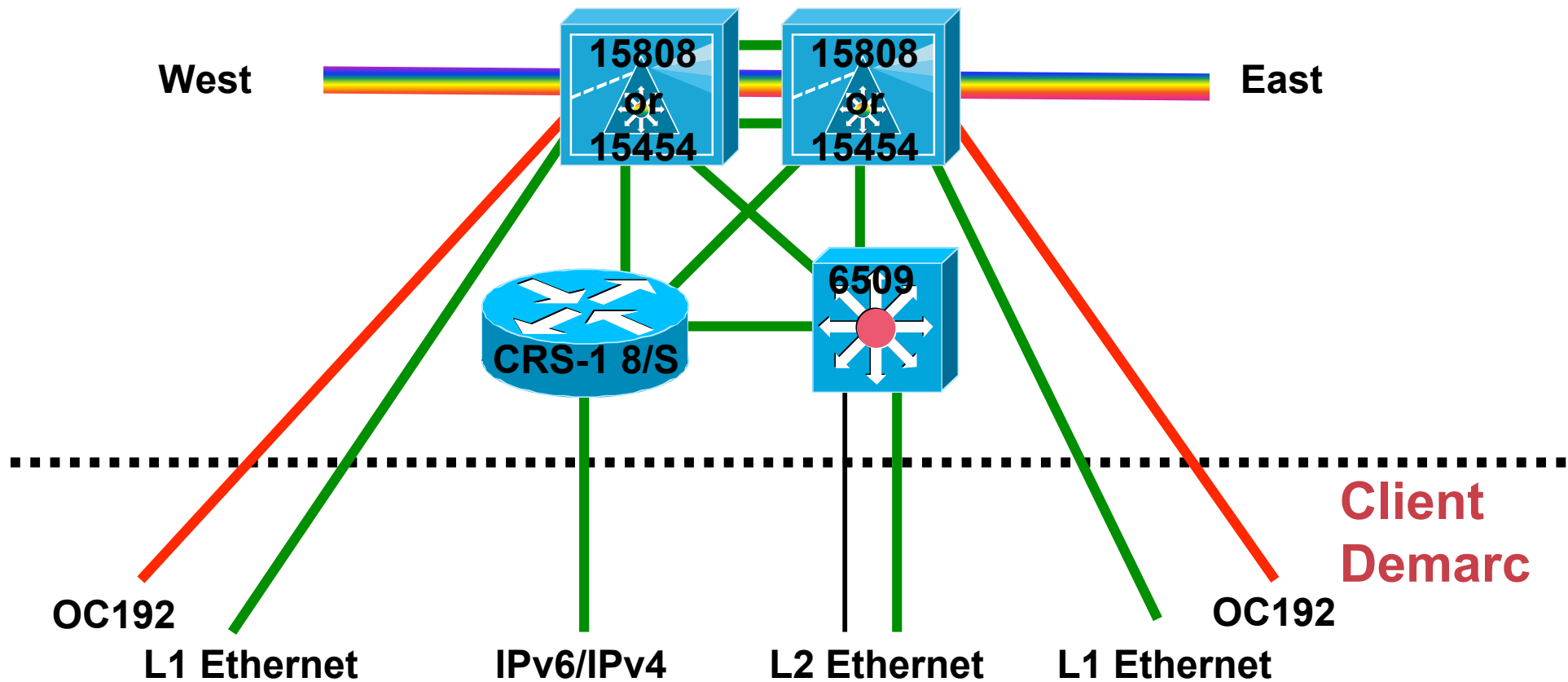
 - L2 10 Gigabit Ethernet Production Backbone**

 - L1 10 Gigabit Ethernet Lambda Backbone**

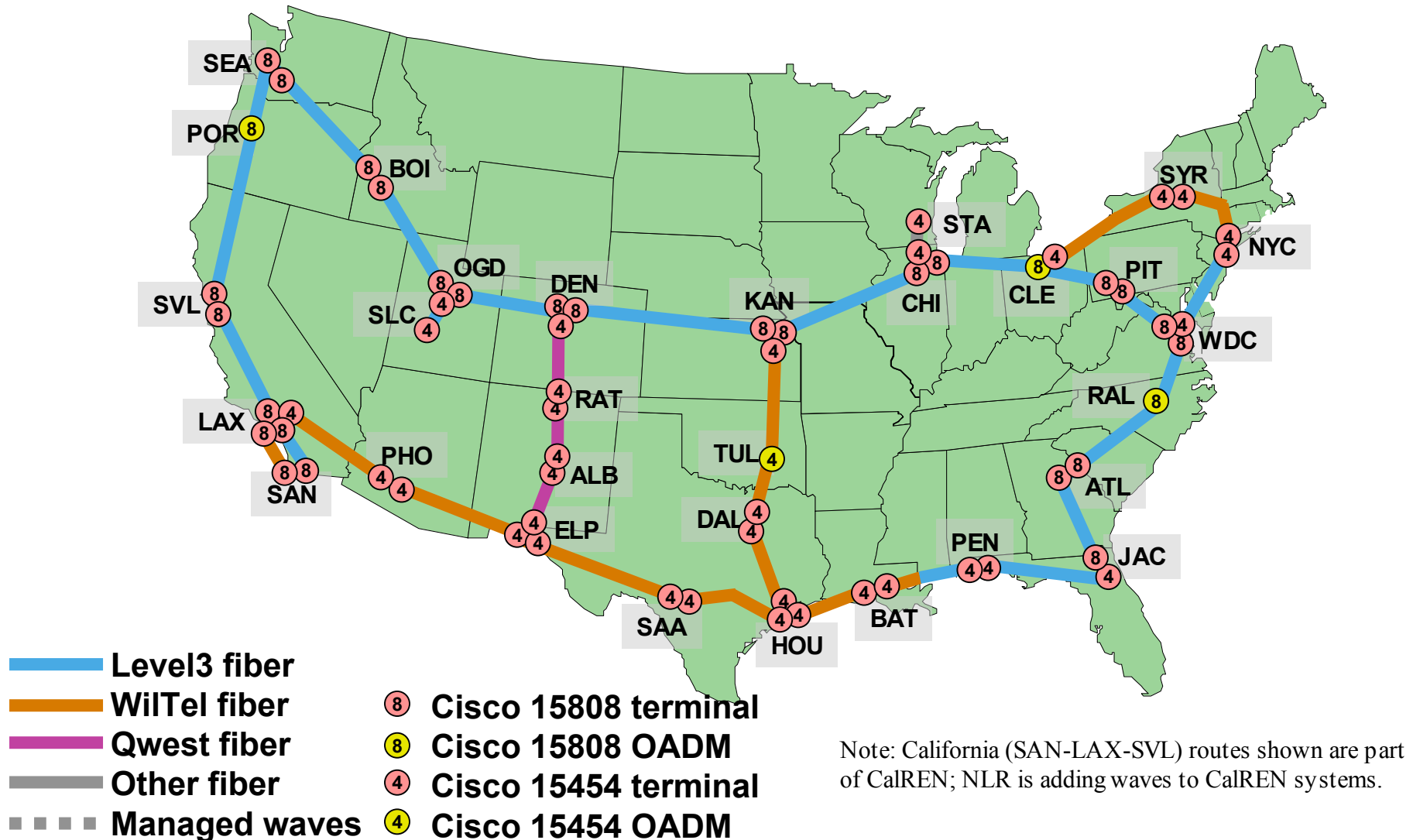
NLR Site Architecture

- STM-64/OC192
- 10 Gigabit Ethernet
- STM-16/OC48
- Gigabit Ethernet
- DWDM Lambdas

Core Site

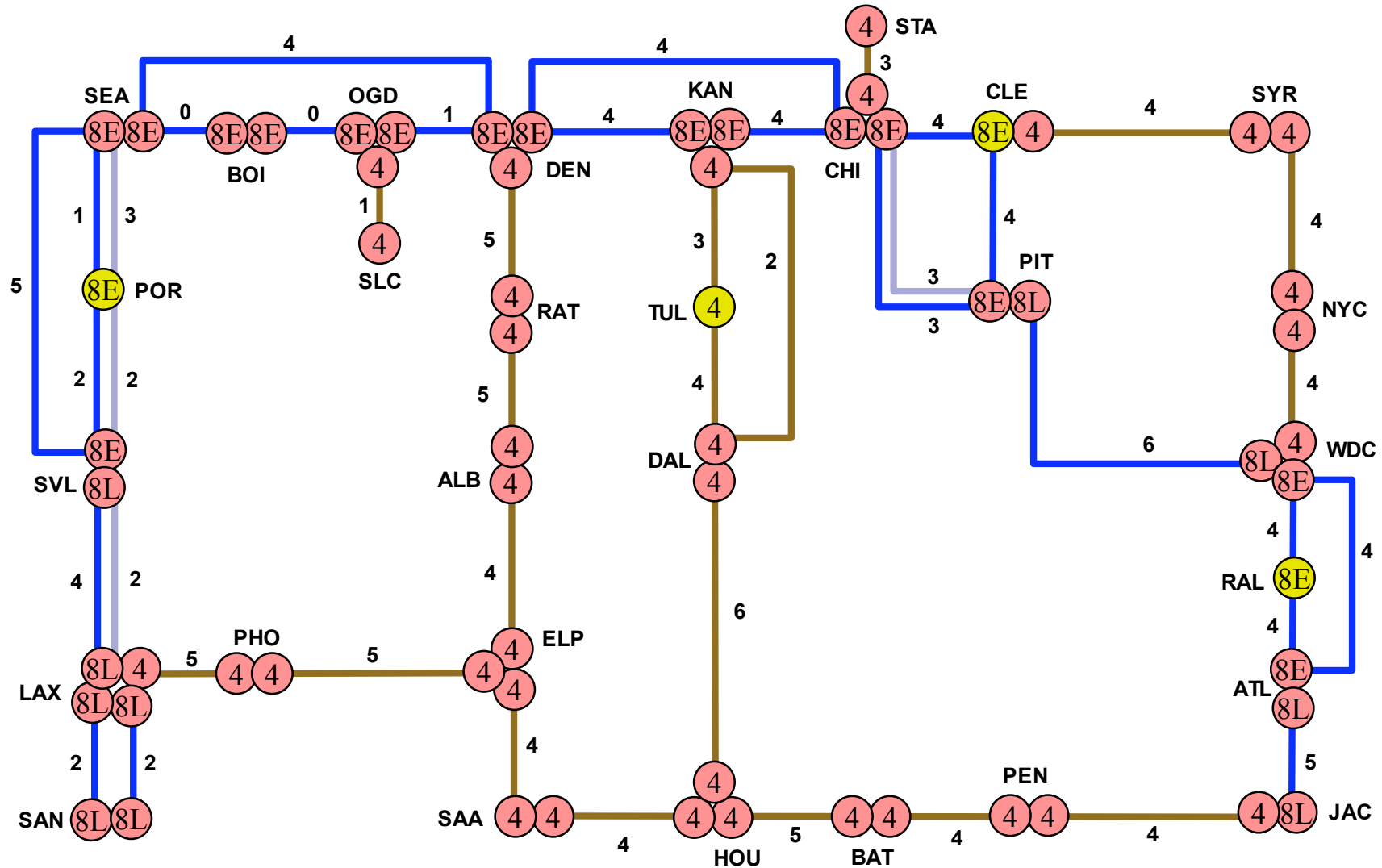


NLR DWDM Backbone



NLR Wavemap

- 10GE wave
- OC192 wave
- 10GE or OC192 wave
- ⋯ 10GE or OC192 managed wave
- 4 15454 Terminal
- 4 15454 OADM
- 8L 15808 LH Terminal
- 8E 15808 ELH Terminal
- 8E 15808 ELH OADM



What is Common in today's NREN's

- **Ethernet is the preferred medium**
 - Gigabit Ethernet is the **lowest** common denominator**
 - 10 Gigabit Ethernet is starting to be deployed at the campus edge**
- **Hybrid Networks are being deployed**
 - L1 Ethernet Circuits - e.g. GLIF**
 - L2 Ethernet Services - e.g. Pacific Wave**
 - L3 IPv6/IPv4 over Ethernet Services - e.g. Grangenet**
- **Production & Research networks are kept away from each other but over a common optical infrastructure**
- **Just deploying an IP Infrastructure is no longer enough for our research communities**
 - “I want a 10 Gigabit Ethernet between my cluster and CERN with nothing in between” - CALTECH High Energy Physicist**
 - I want a Gigabit Ethernet Circuit between my Scanning Electron Microscope in Sydney and the Metalurgist in Melbourne” - GrangeNet Researcher**

Enter the Radio Astronomers with SKA

- **SKA or “Square Kilometer Array”**
 - There will be 1,000,000 square meters of actual collecting dish area
 - The array of antennas will span 3,000km in a number of spiral arcs
 - 50% of the antennas will be within 100km of the apex of the spiral arcs
 - The remaining 50% spanning in a logarithmic series of spiral arcs up to 3000km from the apex
 - Scheduled to “go live” by 2014
 - There are 4 competing site locations for SKA these being:
 - South Africa, Australia, China, Argentina
 - The site winner will be chosen in 2007
 - Have a look at <http://www.skatelescope.org> and <http://www.askac.org>
- **Demonstrators for the SKA are being built now**
- **This research initiative will fundamentally change how we architect, design, implement operate NREN’s and our GRID’s**
- **Yes that is a very bold statement but lets just look at some of “the numbers”.**

Sensor Collection

Each Sensor Gathers:

8 bits of Sampling

2 bits of Polarisation

2 bits of Nynquist Sampling

Samples at the rate of 256 samples / u second

Note this may move to 512 samples / u second

$8 \times 2 * 2 * 256 \times 10E6 = 8.19E9 \text{ bits/s} =$

7.6 Gigabits / second

That is a 10GbE Port required / Sensor

Antenna Collection

Each Antenna has 100 Sensors:

There is a focal array of 10 x 10 Sensors

7.6 Gigabits / Second x 100 =

763 Gigabits / Second / Antenna

That is a switch with:

80 10GbE Ports Input

20 40GbE Ports Output

Station Collection

Each Station has 60 antenna:

There are 60 antenna per station

763 Gigabits / Second / Dish x 60 =

45,766 Gigabits / Second / Station

That is a storage requirement of:

5,722 Gigabytes / second

5.6 Terabytes / second

20 Petabytes / Hour

SKA Collection

The SKA has 130 Stations:

There are 130 Stations

45,766 Gigabits / Second / Station x 130 =

5,950,528 Gigabits / Second

That is a storage requirement of:

743,866 Gigabytes / second

726 Terabytes / second

2,554 Petabytes / Hour

A typical experiment runs for 3 days:

2,554 x 24 x 3 = **183,878 Petabytes of Storage**

Reducing the Data Set Size

A First Level of Correlation at the Dish:

There is a 85% reduction of dataset size

7.6 Gigabits / Second x 100 = 335 Terabytes / Hour * 0.15

50.3 Terabytes Storage / Hour

Reducing the Data Set Size

A Second Level of Correlation at the Station:

There is a 85% reduction of dataset size

$$50.3 * 0.15 =$$

7.5 Terabytes Storage / Hour

That is 543 Terabytes / Experiment / Station

This is a reasonable number for long term storage given that a archival file system is used.

You would need ~ 1 Petabyte of storage per station to run the experiments given you can do phase 1 and phase 2 correlation of the RAW data within 7 hours of it “hitting the sensor”.

What type of NREN do you need ?

It is **NOT a data network problem
it is a Storage Problem!**

A correlator needs to look at the same “wave” as it traverses through the array. Ie It needs to look at the same time stamped data.

Leave the data at the station and have a network SAN Environment built so that a CPU can read the sample files across all 130 stations to carry out the 3rd level of correlation.

This is best done with a software correlator running as a GRID application as the nature is a very parallel process.

You DON'T need to copy the 500 Terabyte file around, you just need to have each CPU doing the correlation for time stamp “X” read its little bit on all 130 sites and then grind away at the Fourier transform.

What type of NREN do you need ?

We need to define a new series of protocols that handle BLOCK transfers of data. I will coin the phrase we need to develop a Hyper Disk Transport Protocol or “hntp”

e.g. “hntp://host/filesystem:block.sector”

The http protocol developed by CERN solved their problems in identifying large files and moving them around.

A hntp protocol is needed to search and identify filesystems in a block mode in a distributed, ordered, secure manner where computing resources can work remotely on the data **WITHOUT MOVING the file itself.**

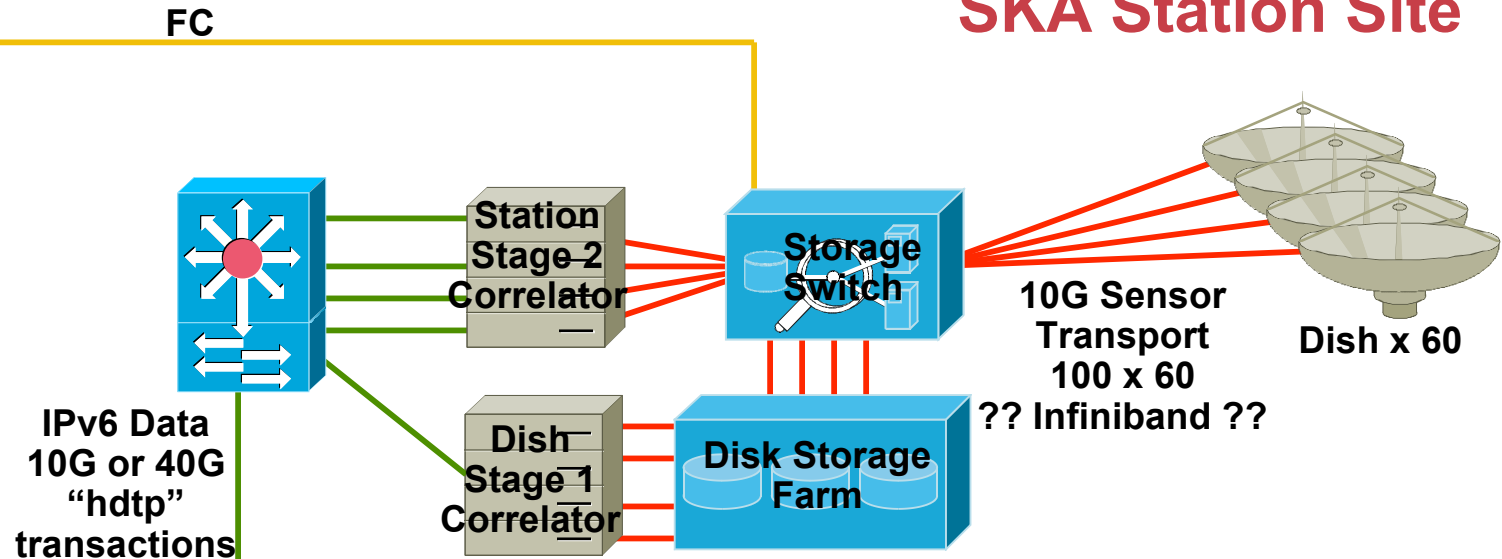
Storage and filesystems need to be block addressable across a network. And NO I AM NOT TALKING ABOUT NFS!

hntp needs to be a bridge between Infiniband, Fibre Channel & IPv6 infrastructures all of which will be critical for SKA to become a reality.

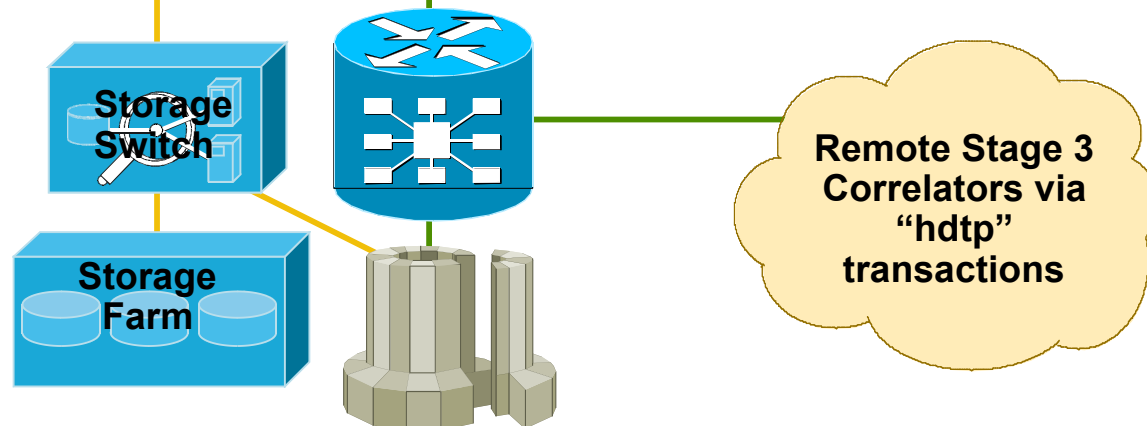
An SKA Architecture

- ??? Infiniband ???
- 10 or 40 Gigabit Ethernet
- 10 or 40 Gigabit FC

SKA Station Site

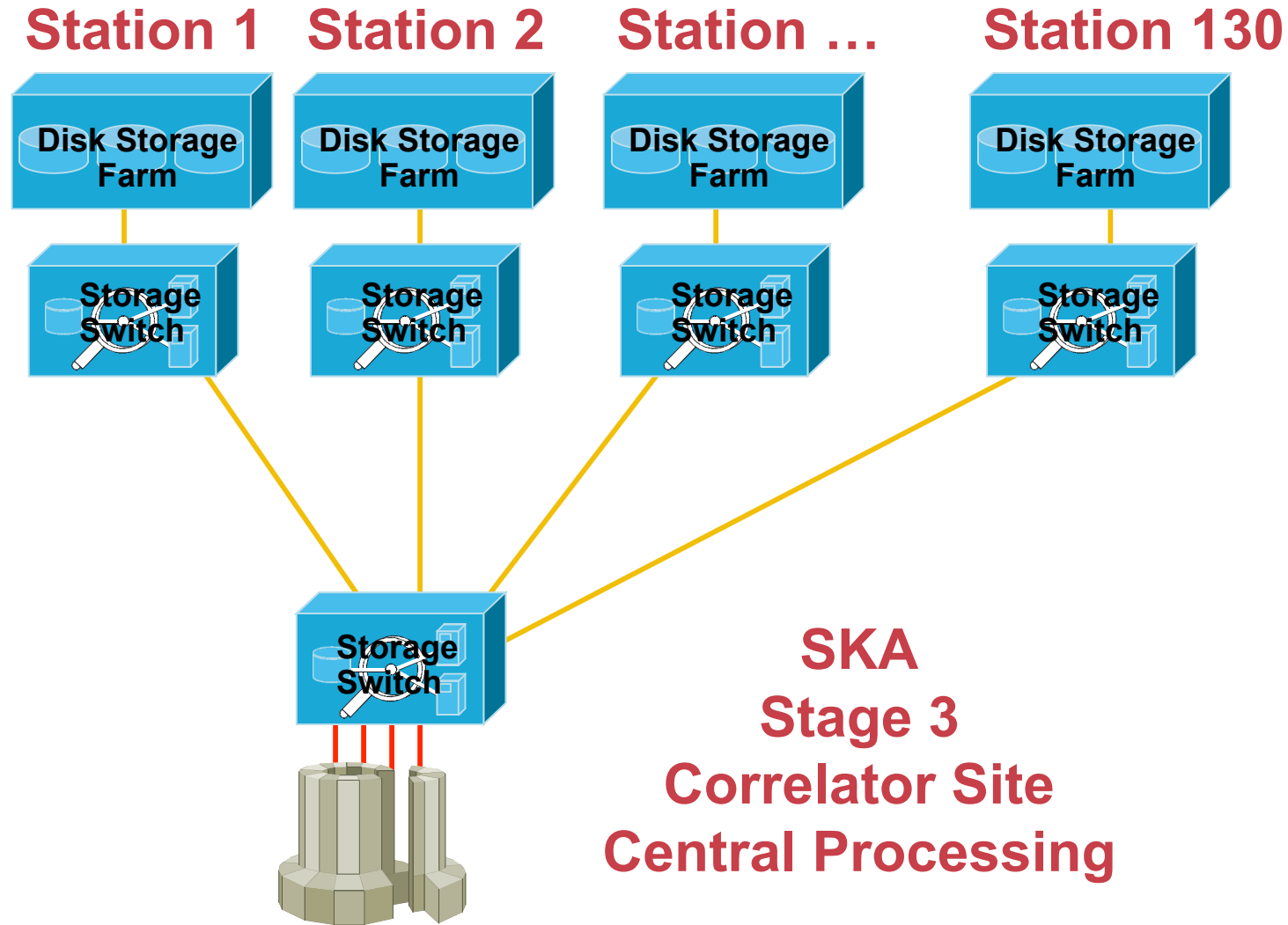


SKA Stage 3 Correlator Site



An SKA Wide Area SAN Architecture

- ??? Infiniband ???
- 10 or 40 Gigabit Ethernet
- 10 or 40 Gigabit FC



Conclusion

- **NREN's need to be considering adding L2 Storage Protocol Transports to their WAN's now.**
- **10G FC with *lots* of Buffer Credits will be a good start**
- **In the longer term a set of protocols need to be developed to route, cache and distribute block mode data between CPU clusters**
- **Data sets will be of a size soon where it will be unaffordable in time, disk space & \$'s to copy files between computing resources.**
- **Storage needs to be treated as a distributed resource not "owned" by a certain computer, super-computer or cluster.**
- **Complete archival file systems with a primary (memory), secondary (disk) & tertiary (tape or DVD) access methods are required. The Plan9 "venti" File System is a good example.**

CISCO SYSTEMS

