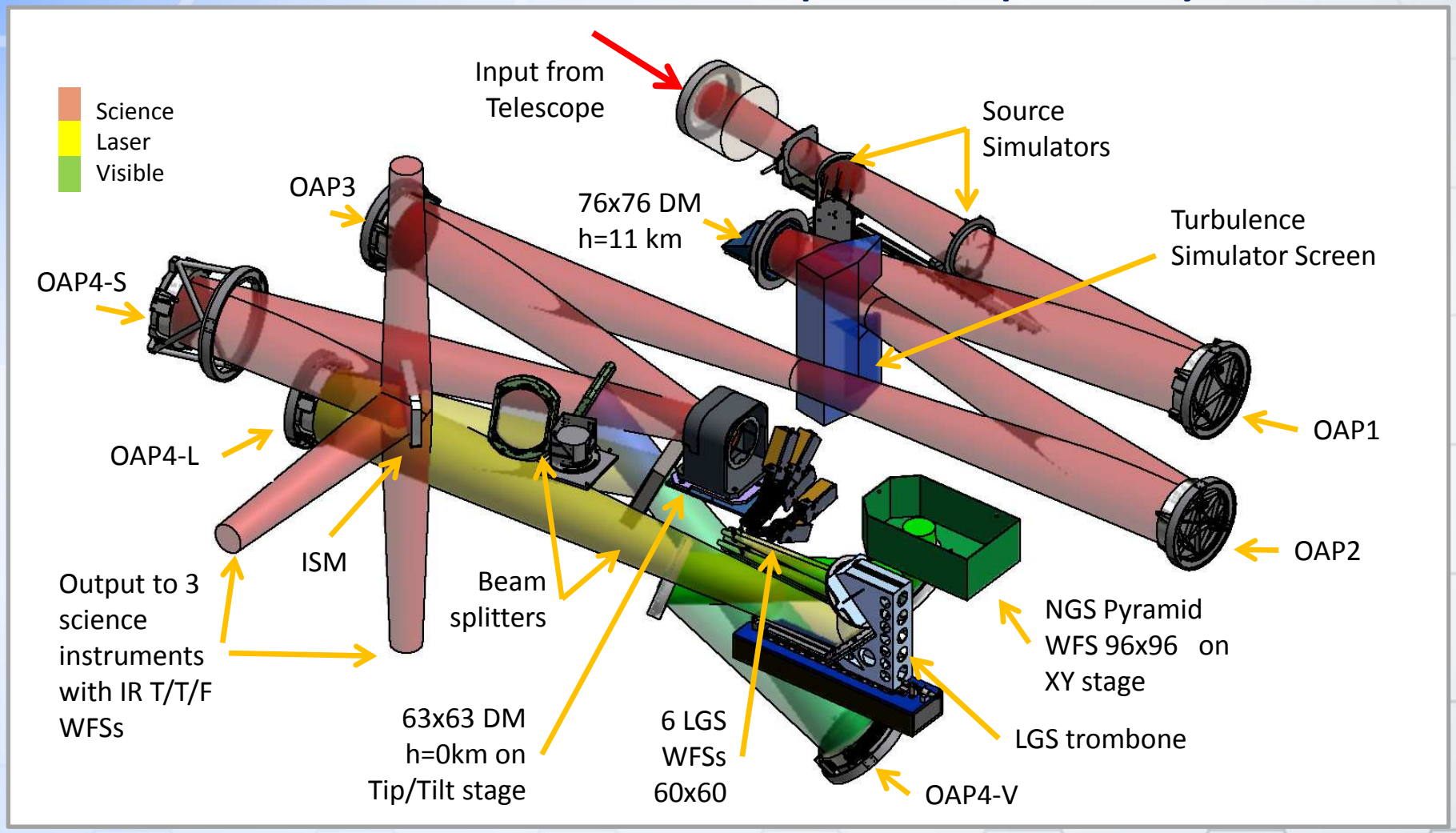
NFIRAOS RTC

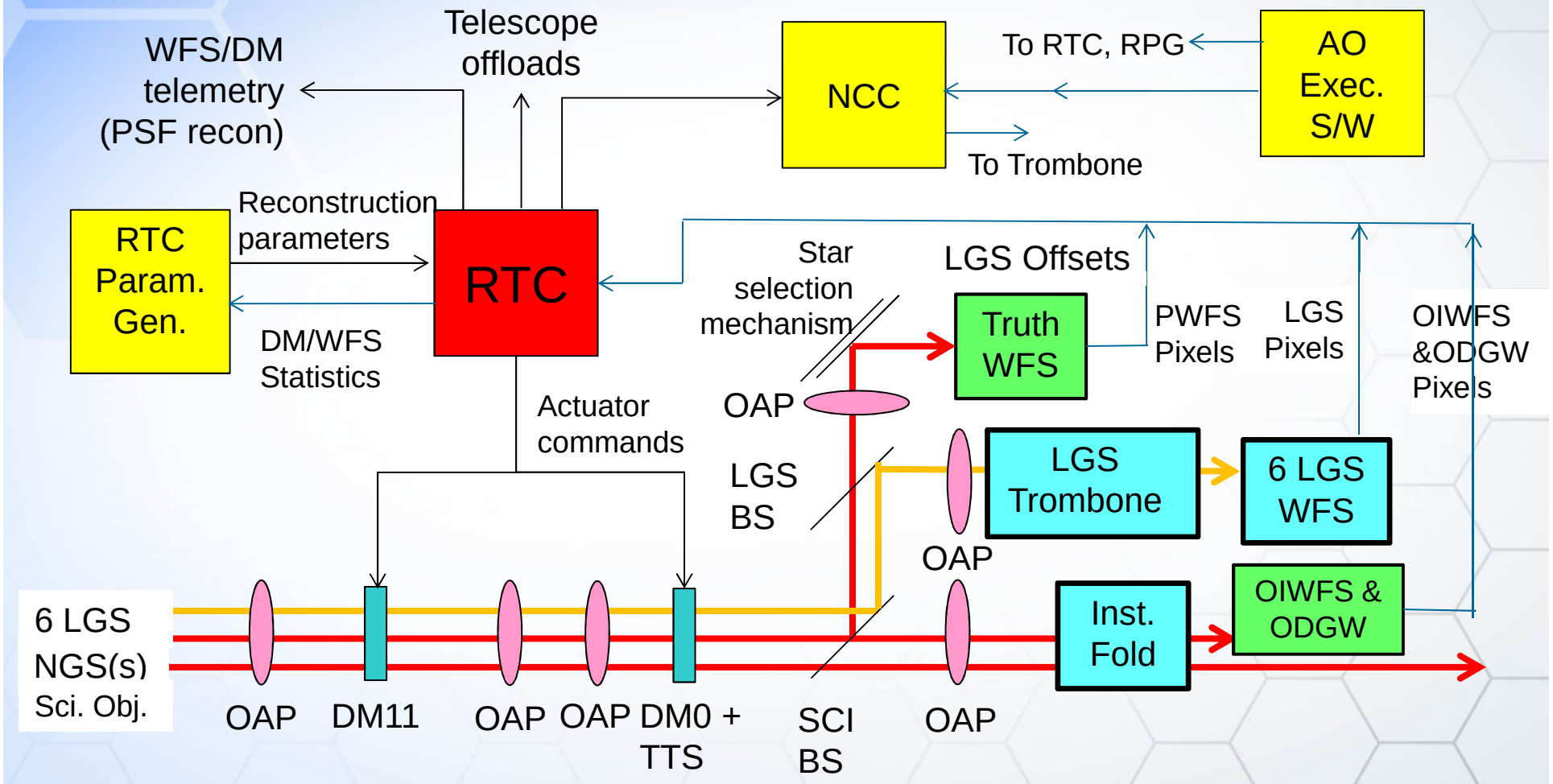
Malcolm Smith
(NRC Herzberg, Victoria, BC)
Paris RTC AO Workshop
December 19, 2016

NFIRAOS

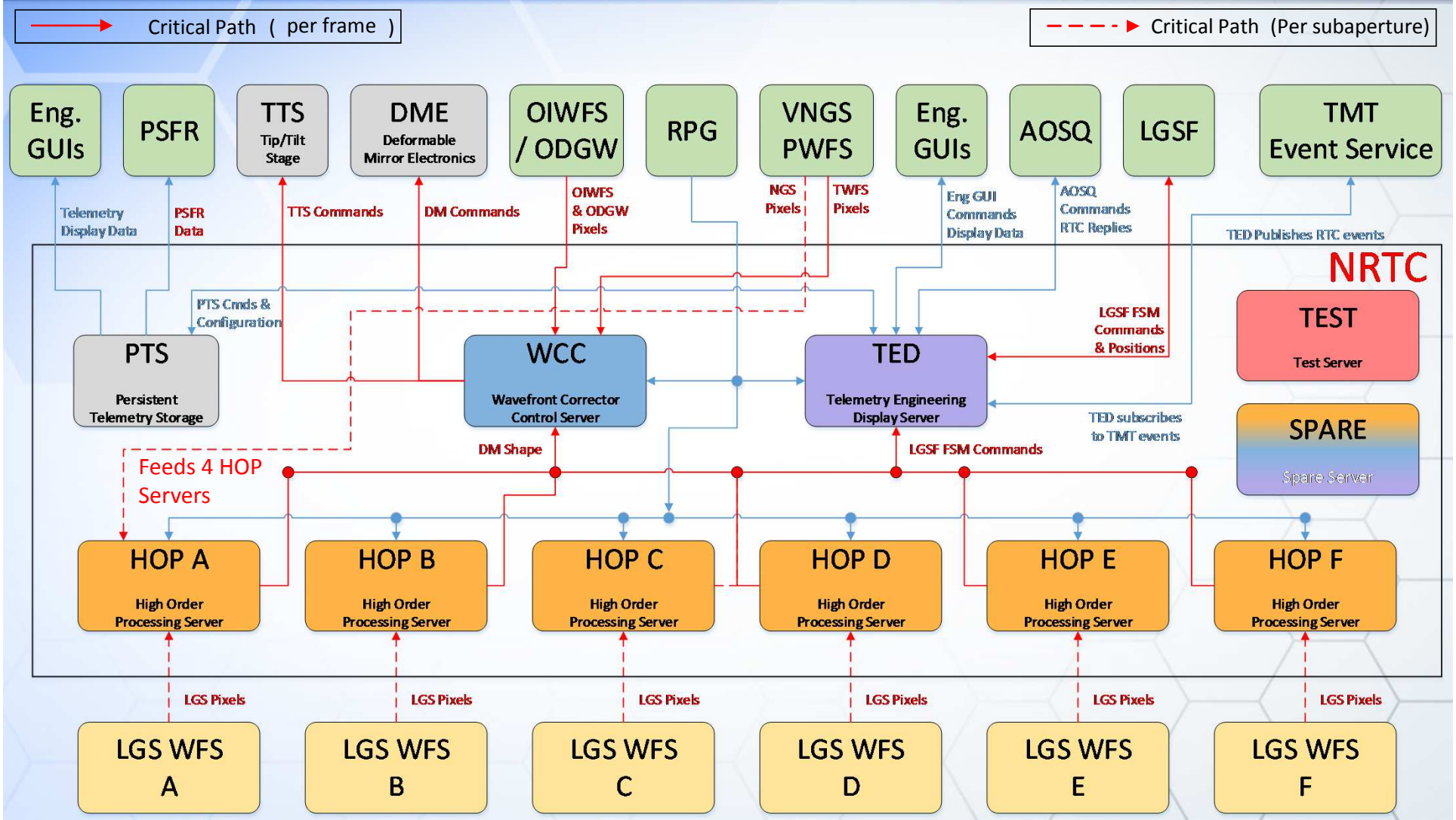
Near Field Infra-Red Adaptive Optics System



NRTC Context



NRTC Overview

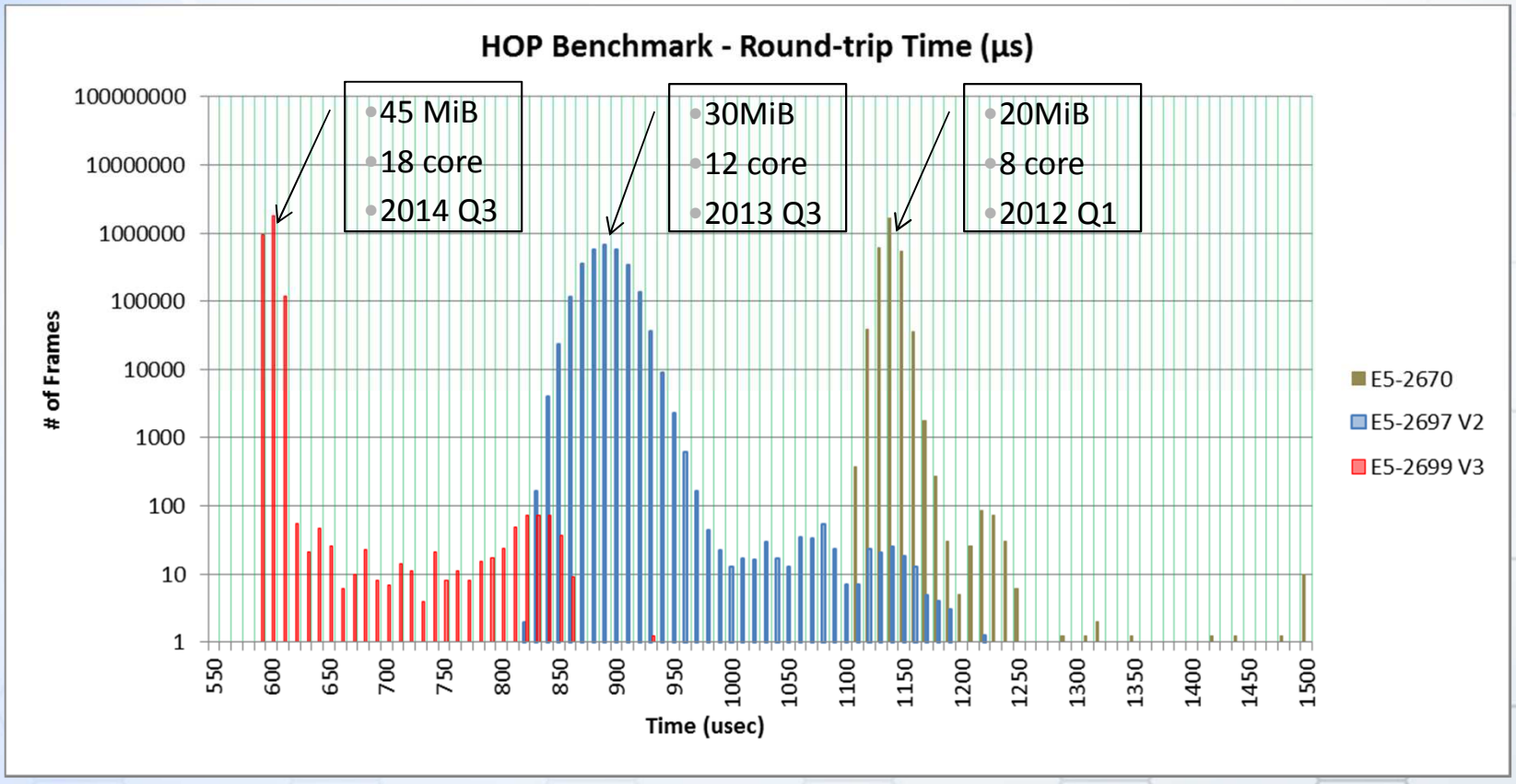


Design Overview

- HOP Servers
  - Read and calibrate high order pixels
  - Compute gradients
  - Perform high order MVM
- WCC Server
  - Read MVM output from 4-6 HOP servers
  - Adjust DM vector (extrapolation, clipping, etc.)
  - Low order processing (e.g. from OIWFS)
- Telemetry Engineering Display Server
- Real-time Telemetry Storage (RTS) Server (was PTS)
- Telemetry Analysis Computer
- Calibration Computer

Early CPU Benchmarks

- Bimodal due to switching control matrix every 10 seconds
- One hour simulated run (360 control matrix updates)
- E5-2600 V4 available 2016 Q1 (up to 55 MiB cache, 22 cores)



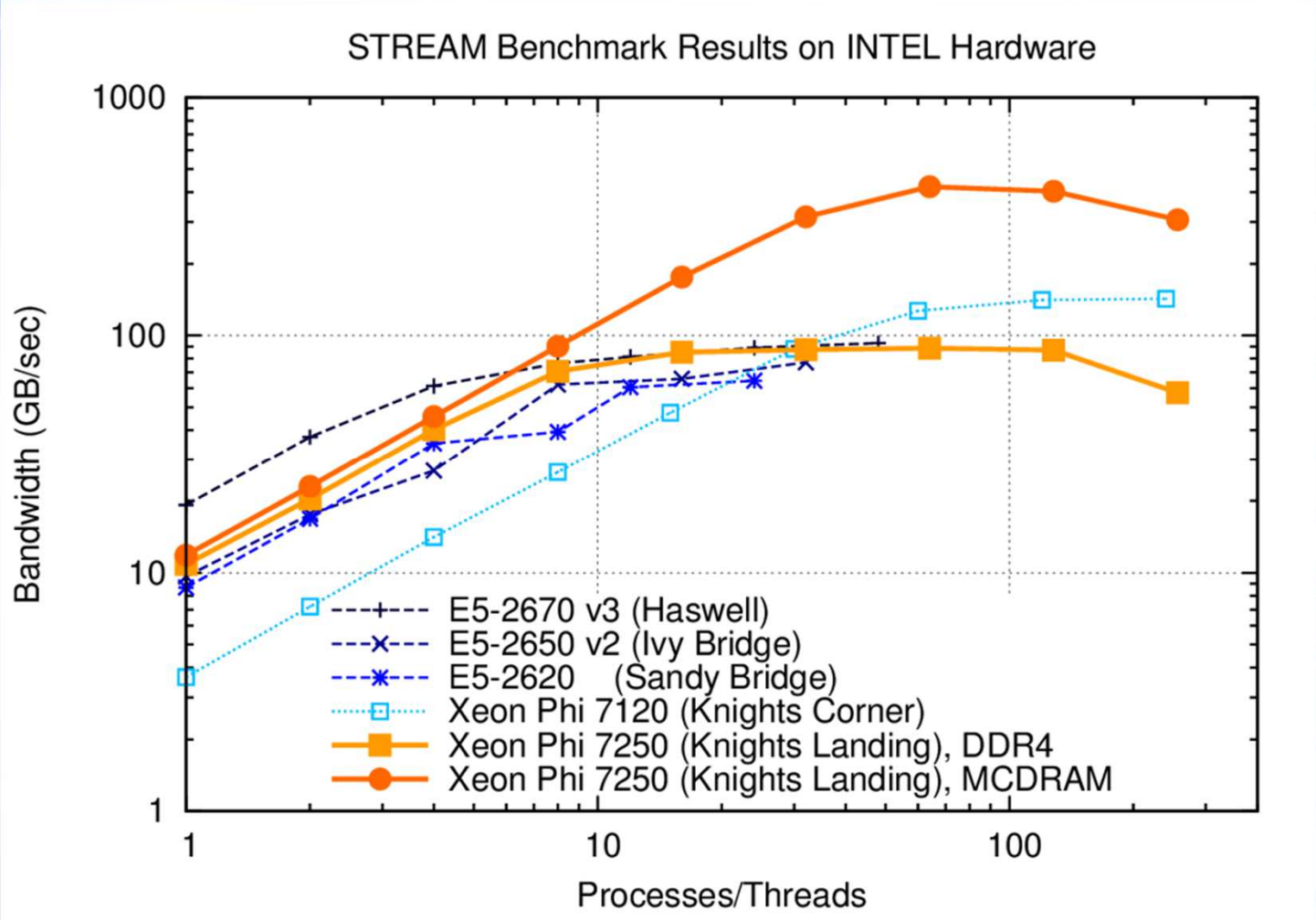
Xeon Phi HOP Servers?

- Can a single Xeon Phi processor can be used in place of four Xeon CPUs?
- Significant cost, space and power savings
  - ~ 30K US for 2U four node Xeon Phi system (2 x $30K)
  - ~ $25K US for 2U quad socket Xeon server (7 x $25K)
  - Prices do not include 10Gb Ethernet, Omni-Path, SSDs
  - Xeon Phi processor ~215 W versus 4 x 135 W
- Borrowed a pre-production KNL 7210 from Intel

  Thanks Intel!
- HOP MVM is bandwidth limited…

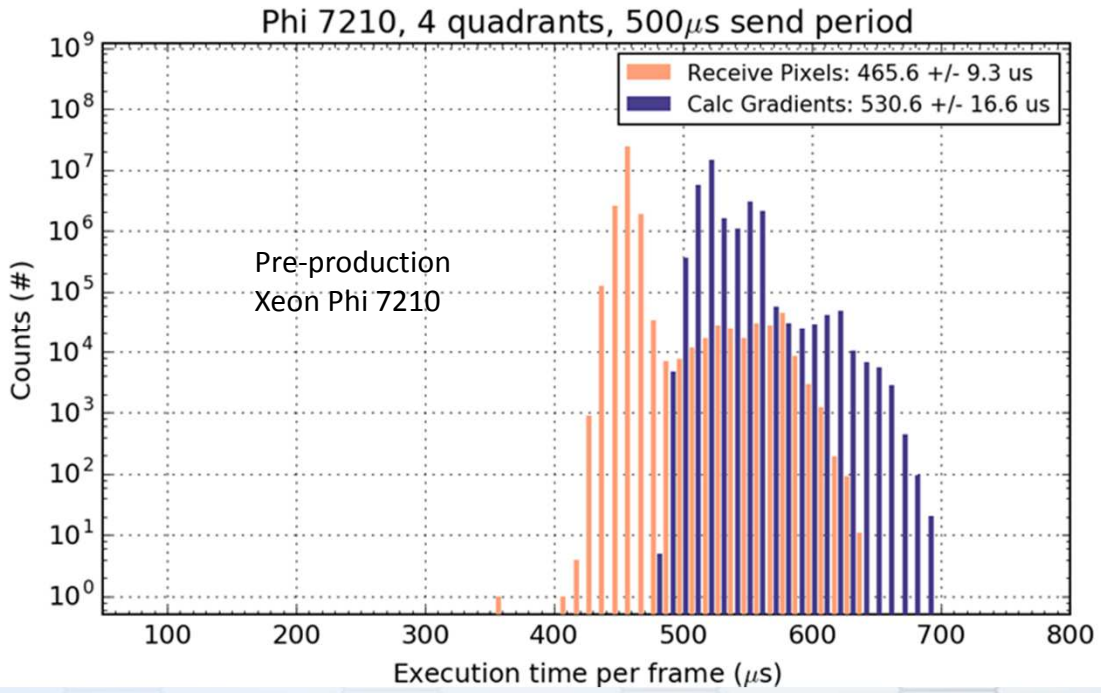
Xeon Phi Memory Bandwidth

- Stream results (& image) by Karl Rupp (www.karlrupp.net)



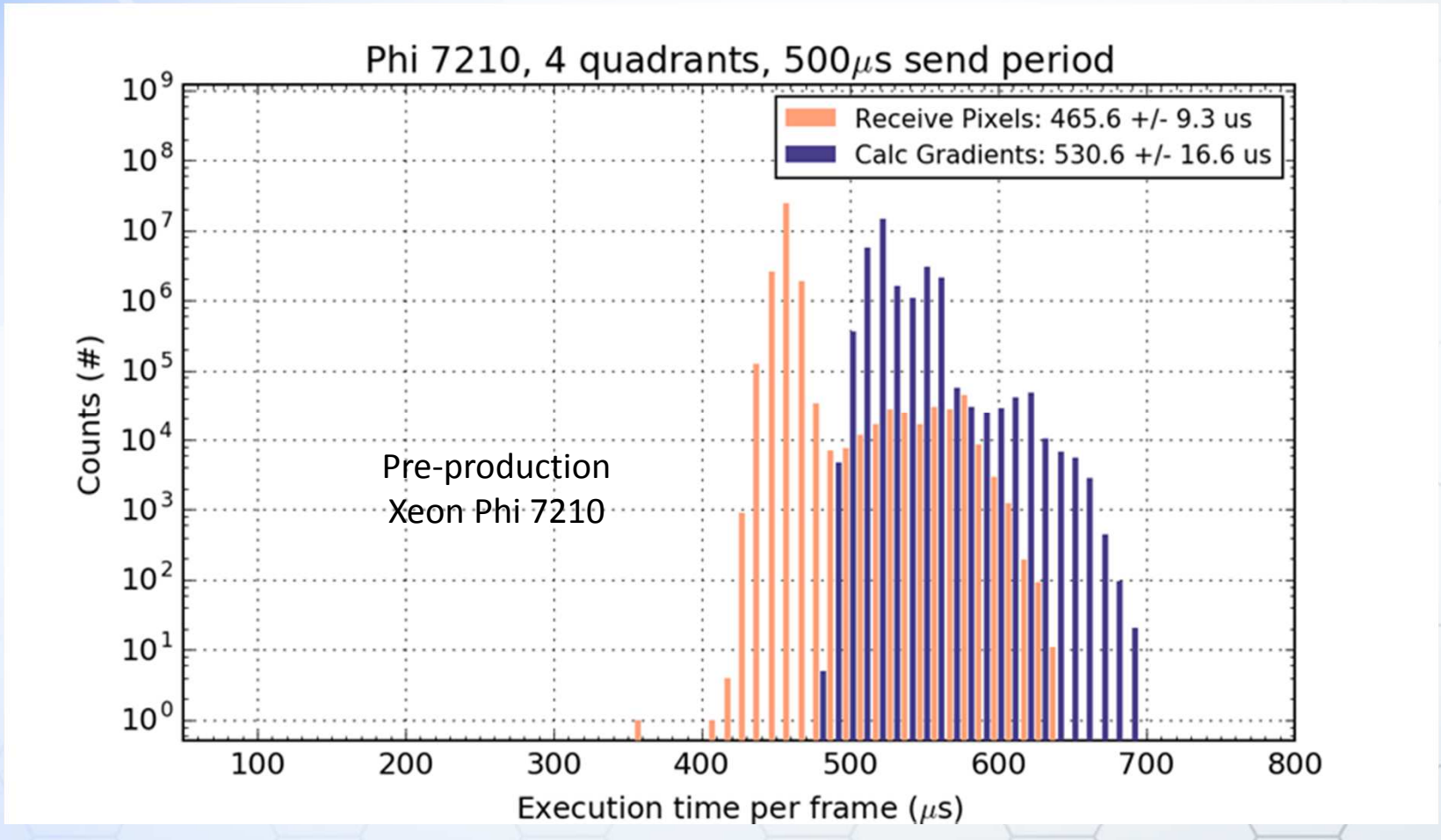
KNL Pixel Processing

- Pixels sent (10 hrs @ 800 Hz) over a 500 µs interval / frame
  - Xeon Phi can keep up but three cores / quadrant required
  - One core for Ethernet interrupts and one core for reading
  - One core for pixel processing (calibration & gradients)
  - Pixel processing usually complete within 600 µs



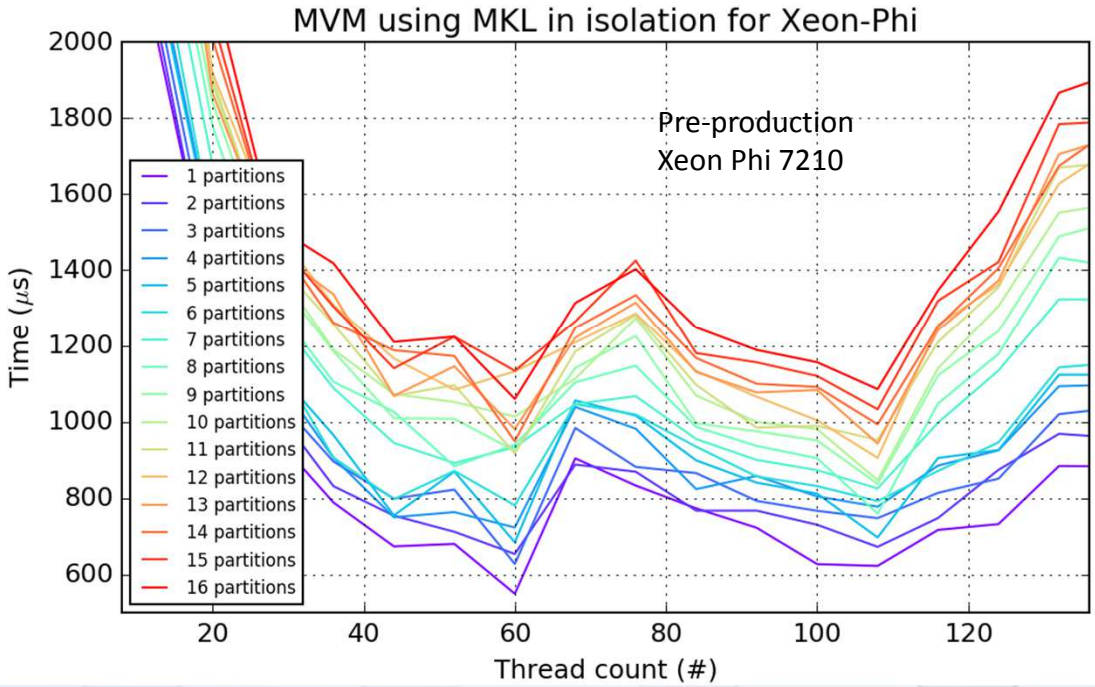
KNL Pixel Processing

- Uses UDP, no datagrams dropped in 10 hours



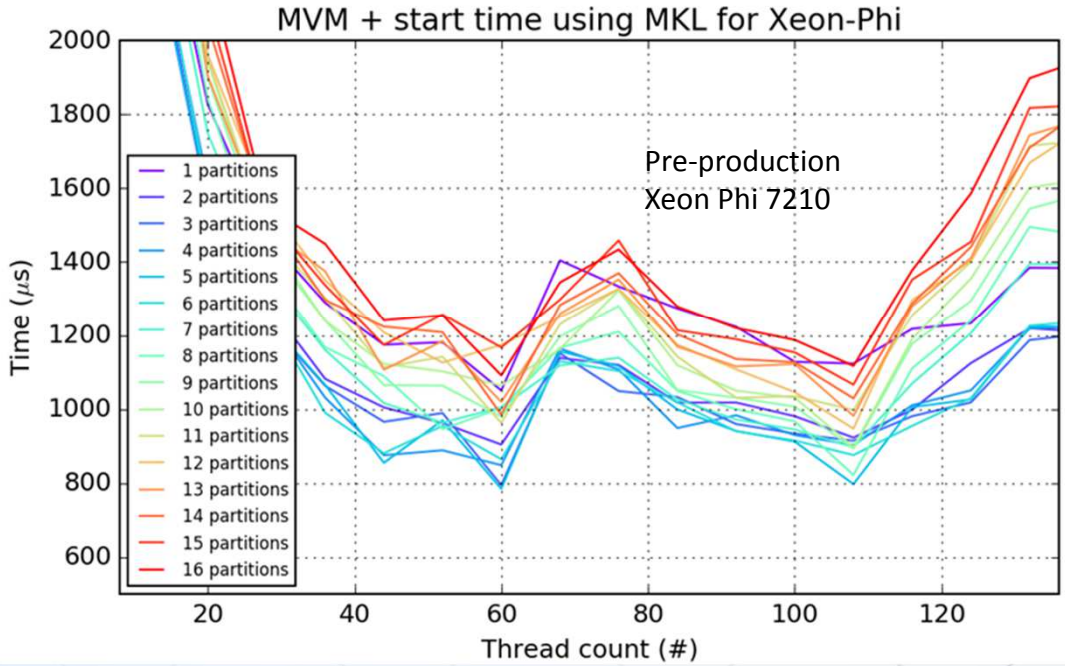
Xeon Phi MKL MVM

- LGS MVM on KNL Phi using Intel Math Kernel Library
- MVM requires less than 600 µs when using 60 threads
- Using ~40 cores, MVM execution time is just under 700 µs
- MKL MVM requires entire input vector



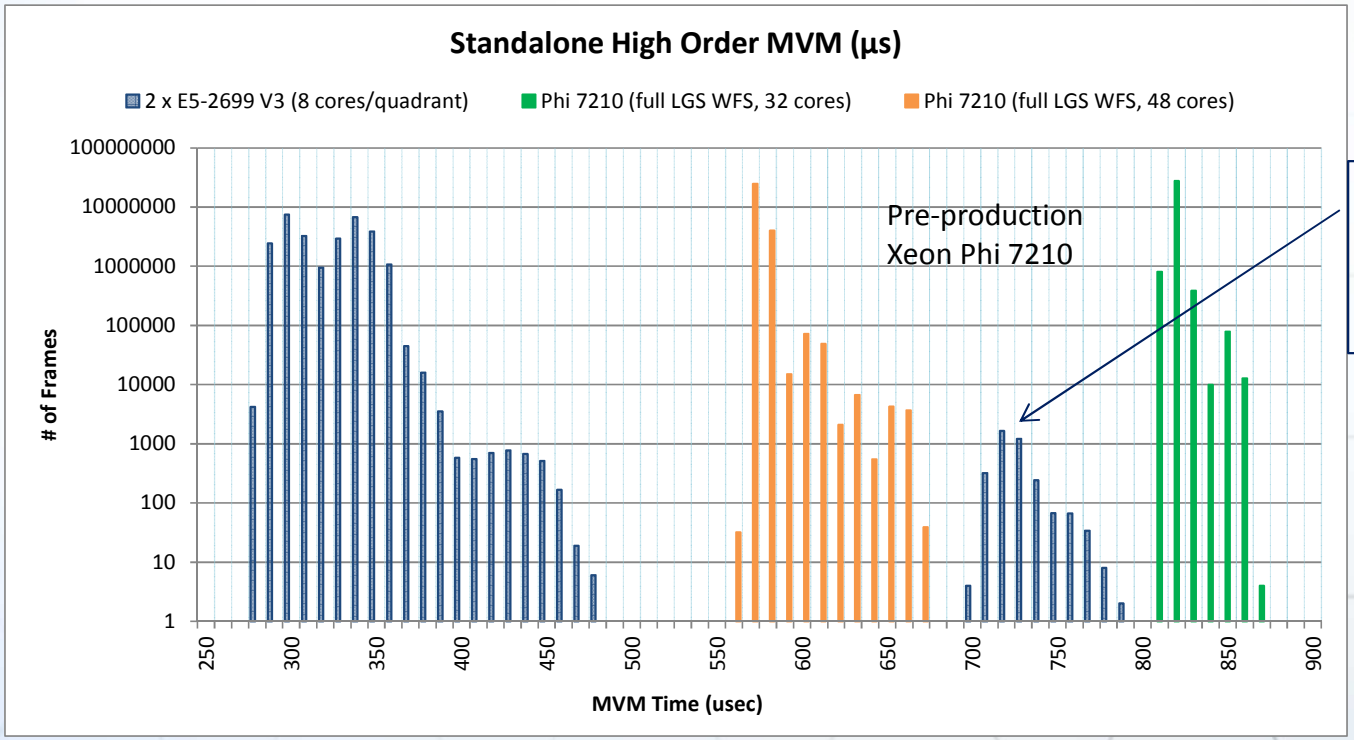
Xeon Phi MKL MVM (2/2)

- Startup delay (500 µs / partitions) added to MVM time.
- Best case requires ~800 µs using 5 partitions (100 µs startup delay)
- Requires 60 cores (does not leave enough for pixel processing)
- Each computation includes overhead due to thread management
  - Normal uses of MKL perform more work to amortize management over



Xeon Phi Custom MVM

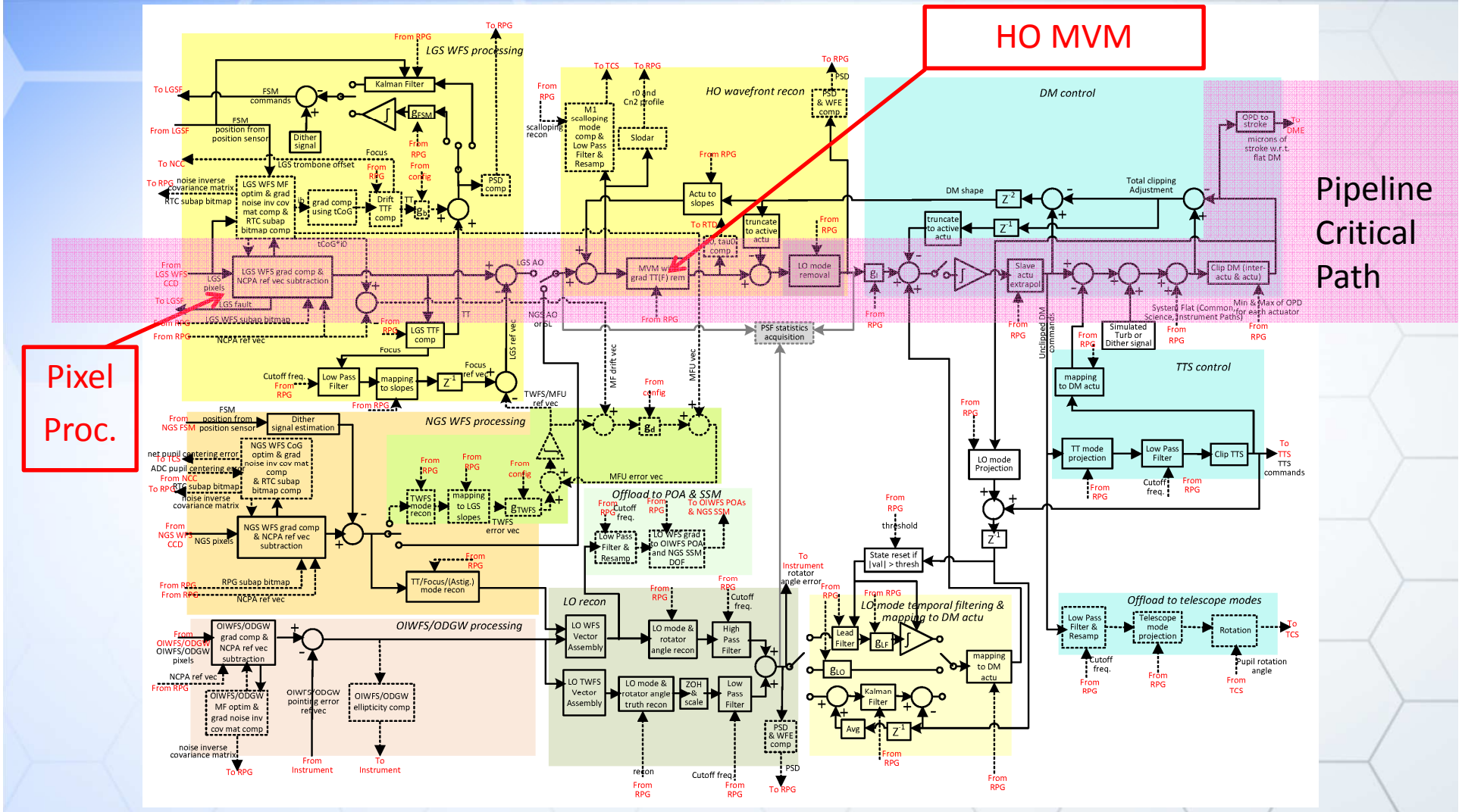
- LGS MVM on KNL Phi and dual socket E5-2699 V3 (2 quadrants)
  - Cached 4 x E5 performance much faster than KNL Phi
  - KNL Phi performance comparable to CPU performance from DDR4
  - Average MVM time for KNL Phi < 600 µs when using 48 cores



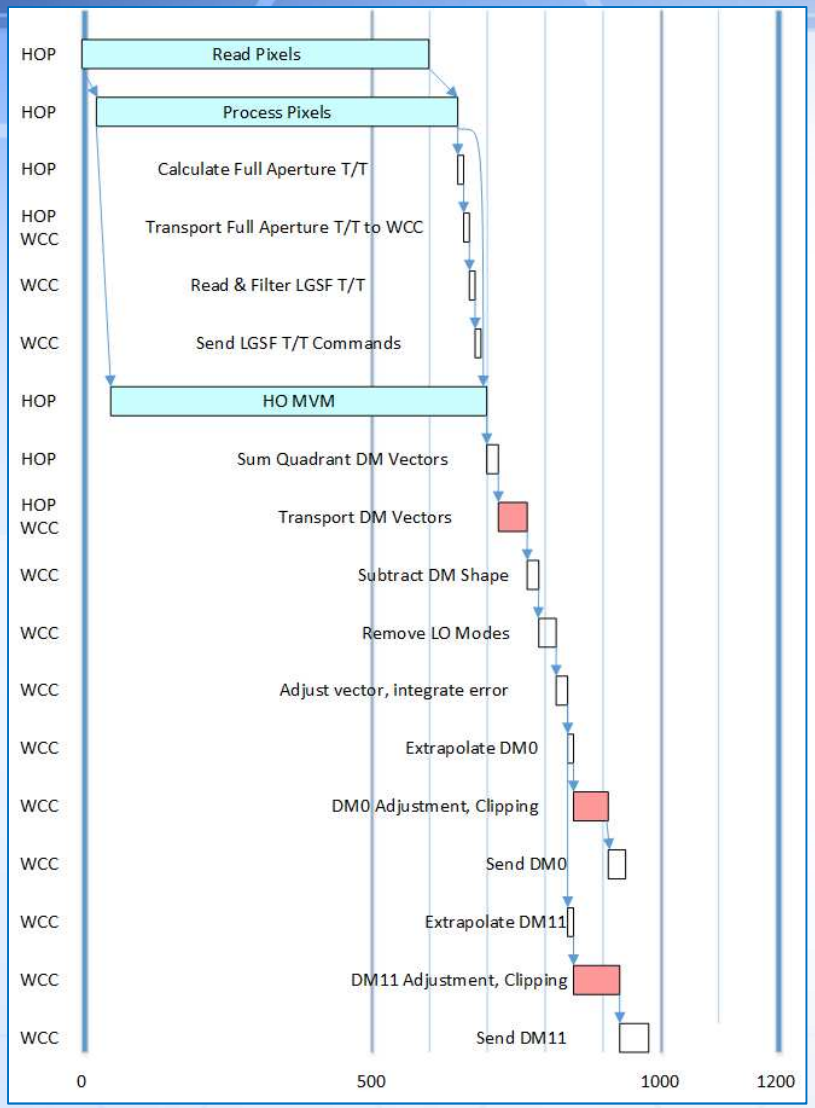
KNL Xeon Phi Performance Notes

- Single Xeon Phi processor **can** be used in place of 4 Xeon CPUs
- Go wide… or go home.
- Xeon CPU is better for **lightly threaded** and/or **scalar** code
  - (e.g. Ethernet interrupts; CPU has higher clock (~1.5X) & smarter cores)
- KNL Xeon Phi provides good performance:
  - If large memory bandwidth is required and MCDRAM is used
    - MVM requires ~3 msec if DDR4 RAM is used
  - If code can be vectorized (use of AVX-512 via pragmas)
  - If code can use many threads (NRTC uses pthreads)
- Newest Intel compiler with explicit KNL code generation provides improved performance and reduced jitter compared to older compiler and gcc.

NRTC Block Diagram (Jan. 2016)



Expected Performance



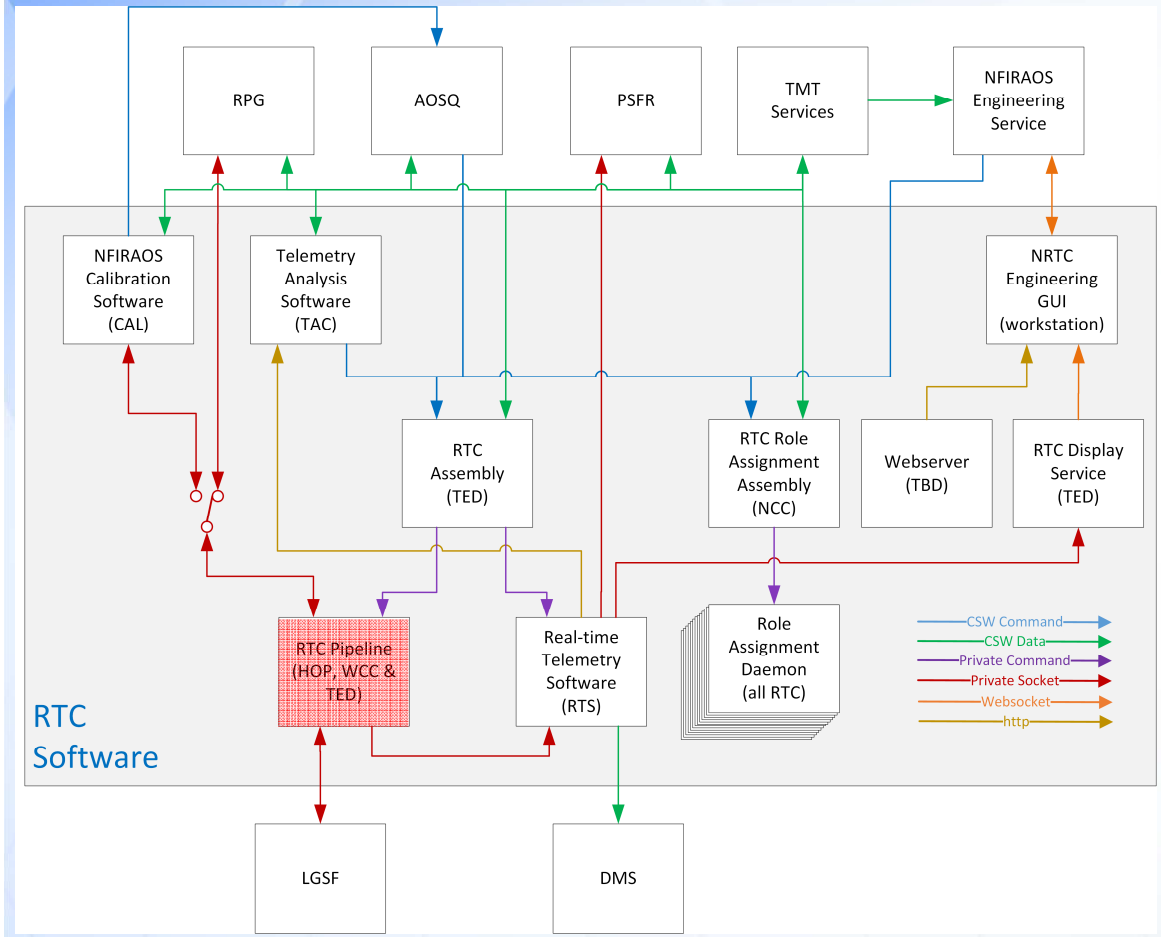
- KNL Phi causes run-time increases:
  - pixel read time,
  - pixel processing,
  - high order MVM.
- Estimates shown based on 1.3 GHz pre-production 7210
  - Baseline 1.4 GHz 7250 uses faster MCDRAM
- HOP server done by 750 µsec
- Time to transport DM vectors assumes Omni-Path fabric (est. ~50 µsec)
- WCC processing assumes benchmarked E5-2643 V4 CPU (6 cores @ 3.4 GHz)
- Average RTC execution time < 1000 µsec

Timings are estimates based on currently specified hardware and benchmarked software. Some optimization (e.g. multi-threaded DM clipping) is possible but will be delayed until build phase.

NRTC Software

- NFIRAOS RTC real-time pipeline
  - only a small part of complete NRTC software
- Many background processing tasks
- NRTC must interface to multiple TMT systems
- NRTC software must enable NFIRAOS calibration
- NRTC software must enable system debugging
  - ~130 TB data stored each night for diagnostic purposes
    - Untagged diagnostic data is deleted prior to observing each night
    - ~60 TB data stored on RTS server
    - Up to ~70TB raw LGS WFS pixels stored on HOP servers
    - Design may be revised to have LGS WFS pixels stored on RTS

RTC Software Components



- RTC has two assemblies
  - RTC Assembly (on TED)
    - Interface to AO Sequencer
  - RTC Role Assignment (NCC)
- RTC Pipeline
  - Implements RTC block diagram
  - Runs on multiple servers
- RTS Software
  - Includes PSFR data (was PTS)
- RTC Display Service
- RTC Test Software
- TAC Software
- CAL Software
- Engineering GUIs
- Role Assignment Daemons

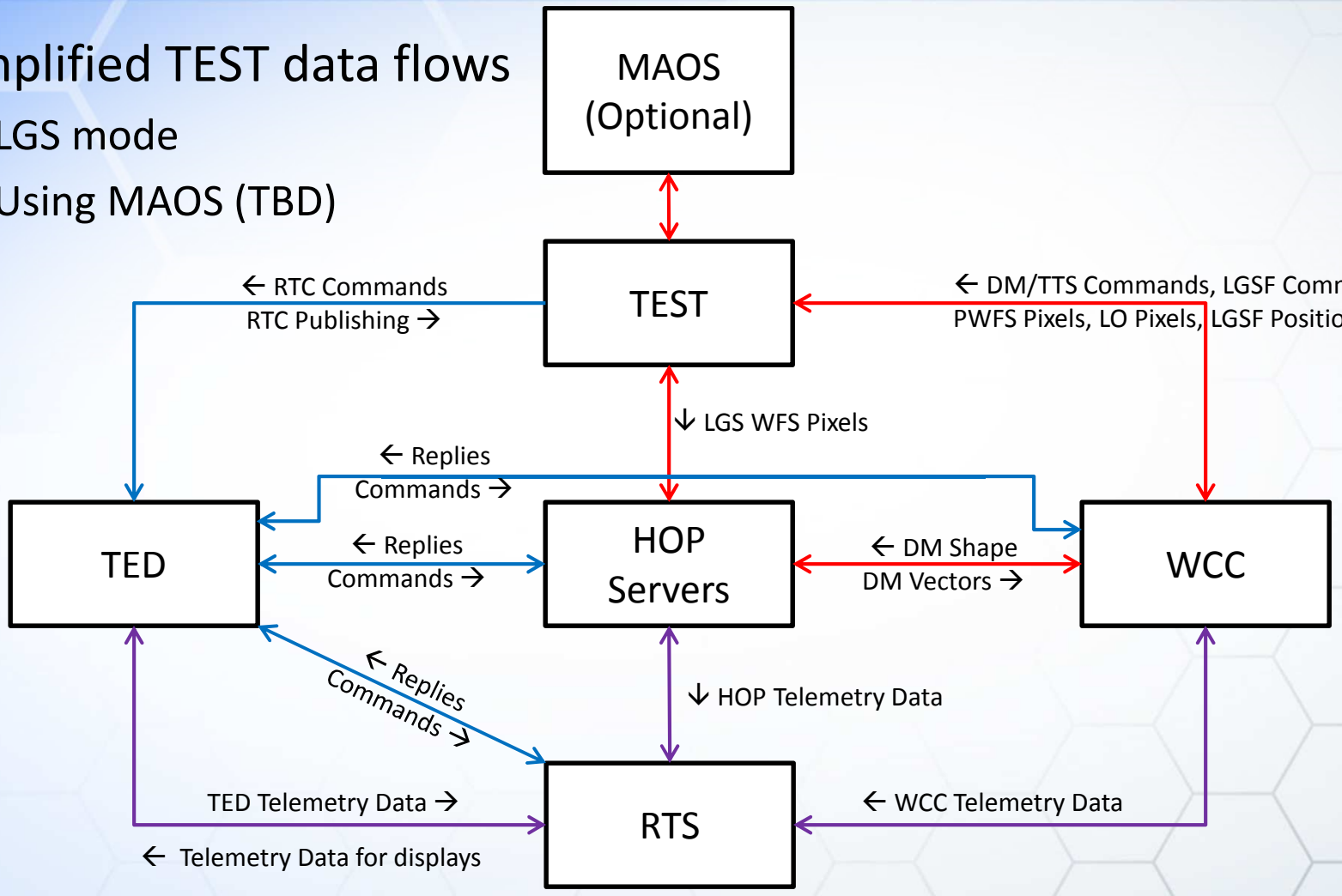
Test Server

- Used to test RTC and verify performance requirements
- Sends the following pixel streams to RTC servers:
  - Up to 6 LGS WFS pixel streams
  - PWFS pixel stream
  - OIWFS / ODGW pixel streams
- Reads DM commands and TTS commands from WCC
- Sends control matrices, etc. to RTC (e.g. simulates RPG)
- Uses two 40Gb Ethernet ports to provide necessary bandwidth
- Reports round trip time from sending of pixels to reading DM vectors and TTS command
- If MAOS is used for closed loop simulation then the interface between MAOS and RTC is strictly via the Test server

Test Server

Simplified TEST data flows

- LGS mode
- Using MAOS (TBD)



NRTC Status

- NFIRAOS RTC final design phase in progress
- Time consuming components of critical path have been prototyped and benchmarked
- Short duration portions of critical path have been estimated
- NRTC can be built using existing technology
  - CPUs could be used for entire system
  - Xeon Phi processors allow significant cost and power savings
- Future work will develop design of remaining NRTC software (i.e. will not concentrate on critical path)

Acknowledgments

The TMT Project gratefully acknowledges the support of the TMT collaborating institutions. They are the Association of Canadian Universities for Research in Astronomy (ACURA), the California Institute of Technology, the University of California, the National Astronomical Observatory of Japan, the National Astronomical Observatories of China and their consortium partners, and the Department of Science and Technology of India and their supported institutes. This work was supported as well by the Gordon and Betty Moore Foundation, the Canada Foundation for Innovation, the Ontario Ministry of Research and Innovation, the National Research Council of Canada, the Natural Sciences and Engineering Research Council of Canada, the British Columbia Knowledge Development Fund, the Association of Universities for Research in Astronomy (AURA) and the U.S. National Science Foundation.

NRC gratefully acknowledges the generous support of the Intel Corporation for providing access to Xeon Phi x200 and Xeon E7 V3 (Dell R930) hardware and current Intel Compiler Suite to allow our benchmarking to continue in a timely manner.

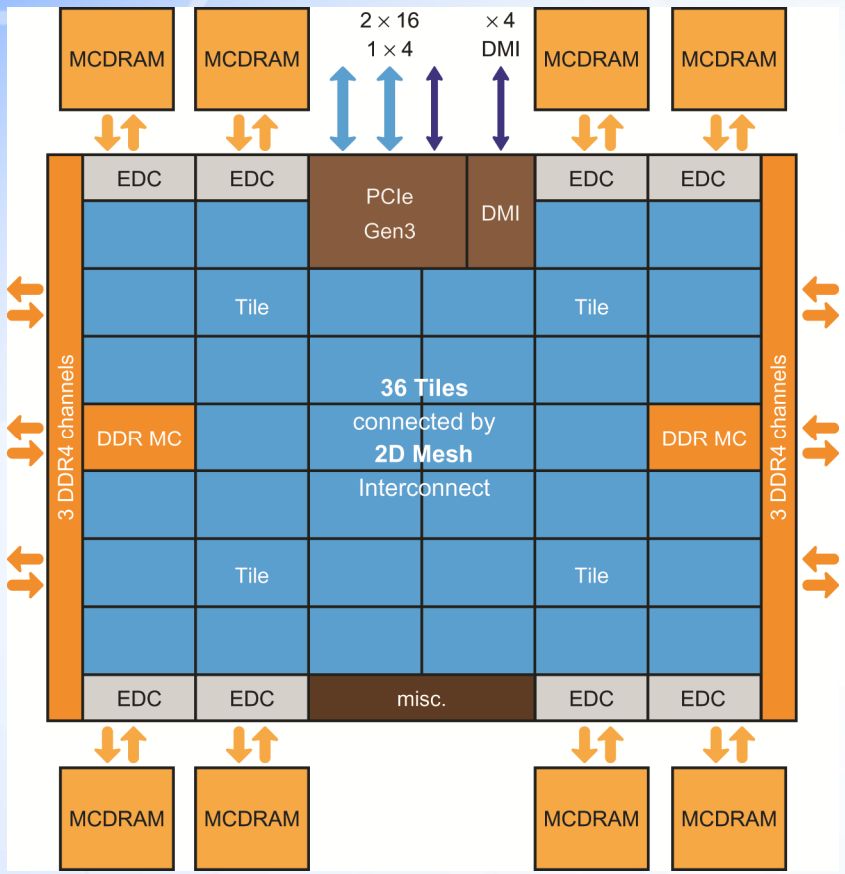
Backup Slides

- Post-PDR Changes
- KNL Phi architecture details
- OS tools (assigning cores, etc.)
- I/O Considerations
- NGS mode, # HOP servers required
- Real-Time Telemetry Storage
- Telemetry Analysis Computer
- Engineering GUIs
- Calibration

Post-PDR Design Changes

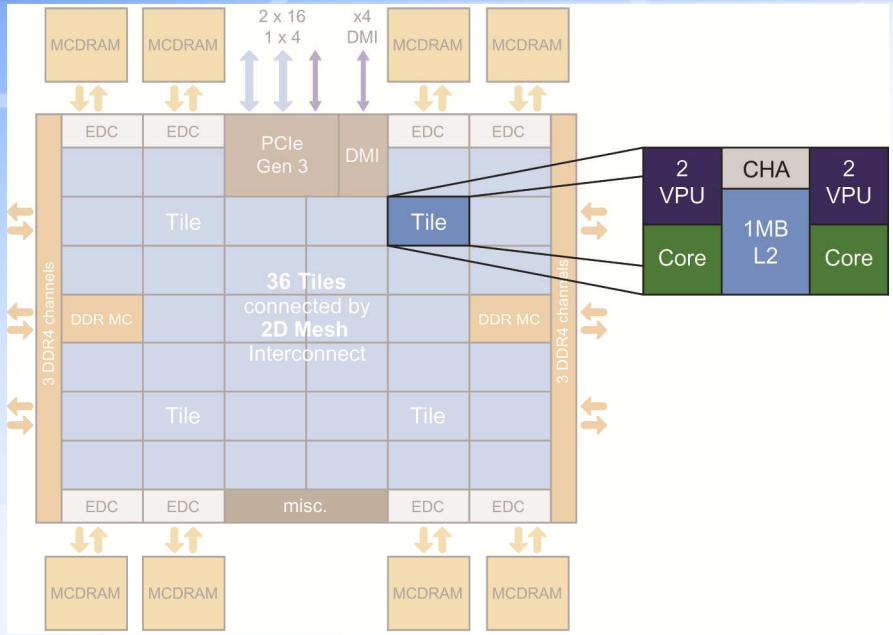
- Identical servers are no longer a goal for design
  - HOP servers now use KNL Xeon Phi processors
  - Other servers are dual socket E5 V4.
    - Dual socket servers can likely standardize on base functionality
    - Differences in PCI-e adapters and SSD capacity in E5 servers
- NGS mode now uses both deformable mirrors
- Telemetry data is (mostly) centralized (on RTS)
  - RTS will contain extra SSDs and RAID controllers
- Internal RTC network traffic will use Omni-Path
  - Significantly simplifies Ethernet network topology

KNL Xeon Phi Architecture



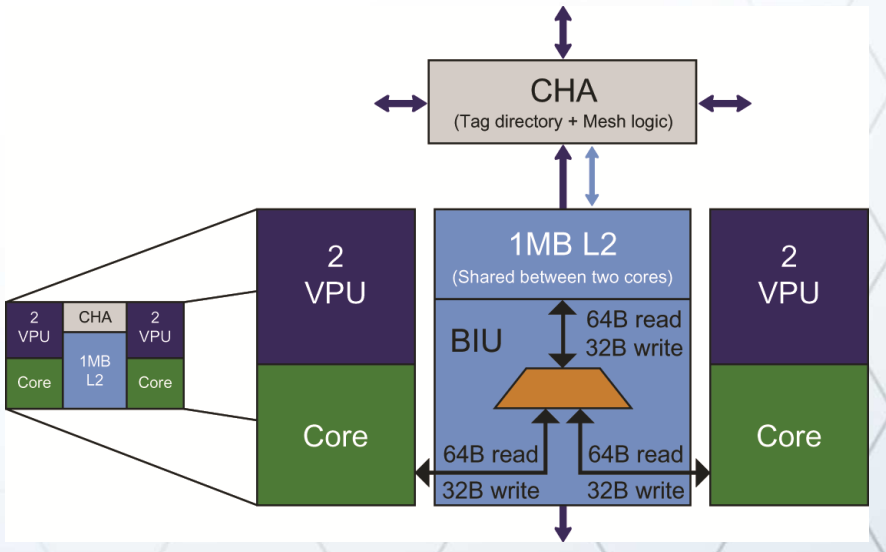
- KNL Phi die contains 38 tiles
  - 2, 4 or 6 tiles are disabled
  - 72, 68 or 64 cores
  - Tiles are connected in 2D mesh and have coherent caches
  - 8 MCDRAM interfaces (2 per quadrant)
  - PCIe connected to first quadrant
  - Sub-Numa Clustering (SNC4) mode makes processor appear as 4 processors & 8 NUMA regions

KNL Xeon Phi Architecture (2)



- Each tile contains two cores
- Each core contains two vector units
- Cores share L2 cache (compete for L2 resources)

- HOP servers not limited by floating point computation
- 32 S.P. floating point ops (FMA) per cycle / VPU
- Model 7250 ~ 6 Tflops (S.P.)
- Entire RTC requires < 1 TFlop



OS Considerations

- Linux with real-time kernel and utilities
  - tuna
- Isolate CPU cores (exclude from general scheduler)
- Assign cores to required interrupts (e.g. 10GbE)

I/O Considerations

- Network interrupt coalescing
- Disk throughput
- Latency vs throughput

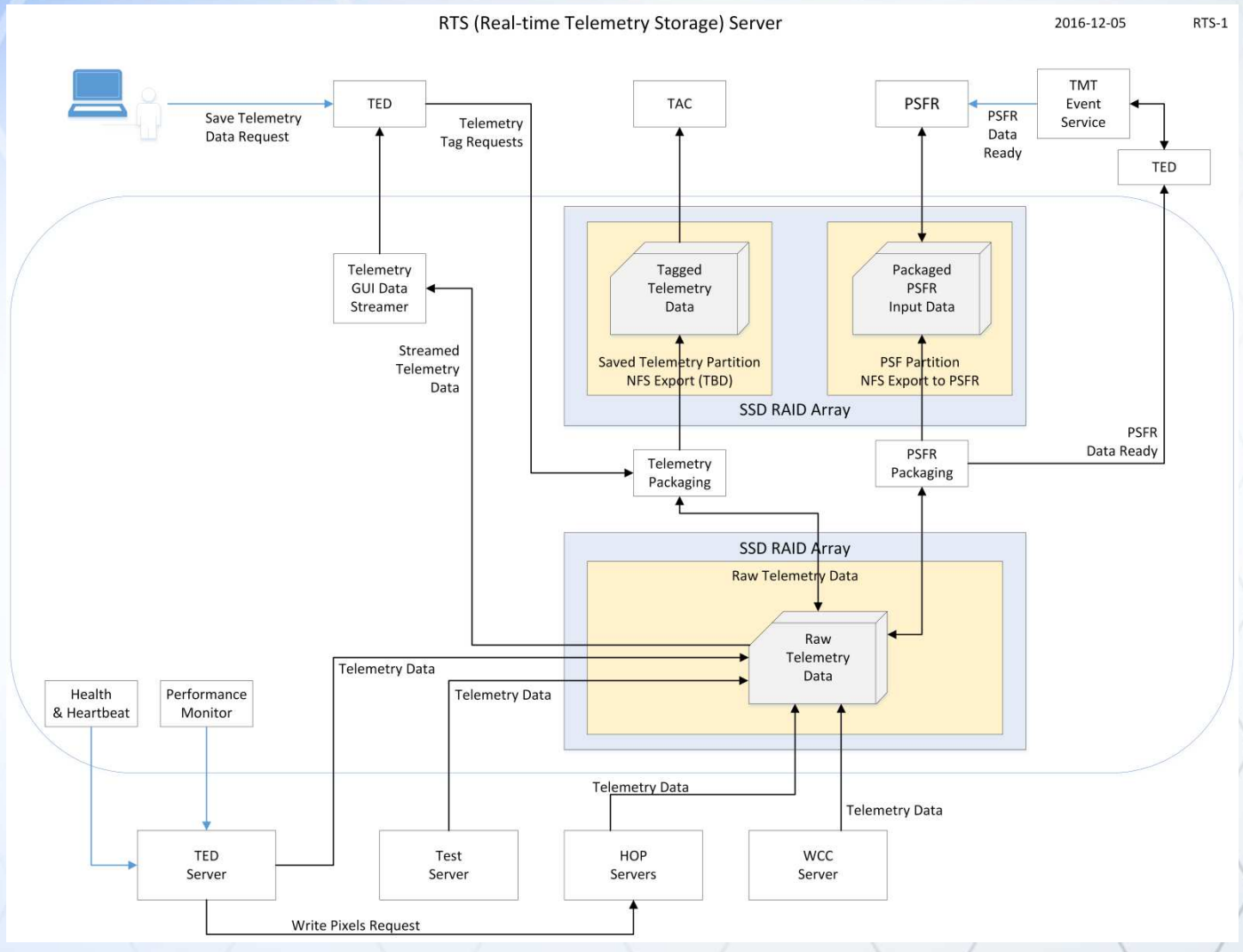
NGS HOP Servers

- HOP servers send DM vectors within 750 µs of frame start
- ~7238 PWFS subapertures ($\pi\ 48^2$ pixels per quadrant)
- PWFS mode has ~14476 gradients (vs 5968 / LGS WFS)
- MVM delayed by ~250 µs
  - Gradients not computed until second half of PWFS readout
- Phi based HOP server computes LGS MVM in ~600 µs
- NGS MVM would require (14476 / 5968) x 600 µs ~ 1450 µs
- Require: 250 µs + 1450 µs / (# HOP) < 750 µs
- Solving for # HOP > 2.9
- Number of HOP servers used for NGS mode = 4

Real-time Telemetry Storage

- LGS WFS pixels stored on HOP servers
- Other telemetry data stored on dedicated RTS server
  - RTS will include multiple RAID controllers
  - RTS will include ~60 TB (TBC) SSD storage
- Currently TBD whether data is written to disk by dedicated threads on RTS or by exporting as NFS
- RTS may store all 130 TB of data
  - Subject to further testing and SSD capacities

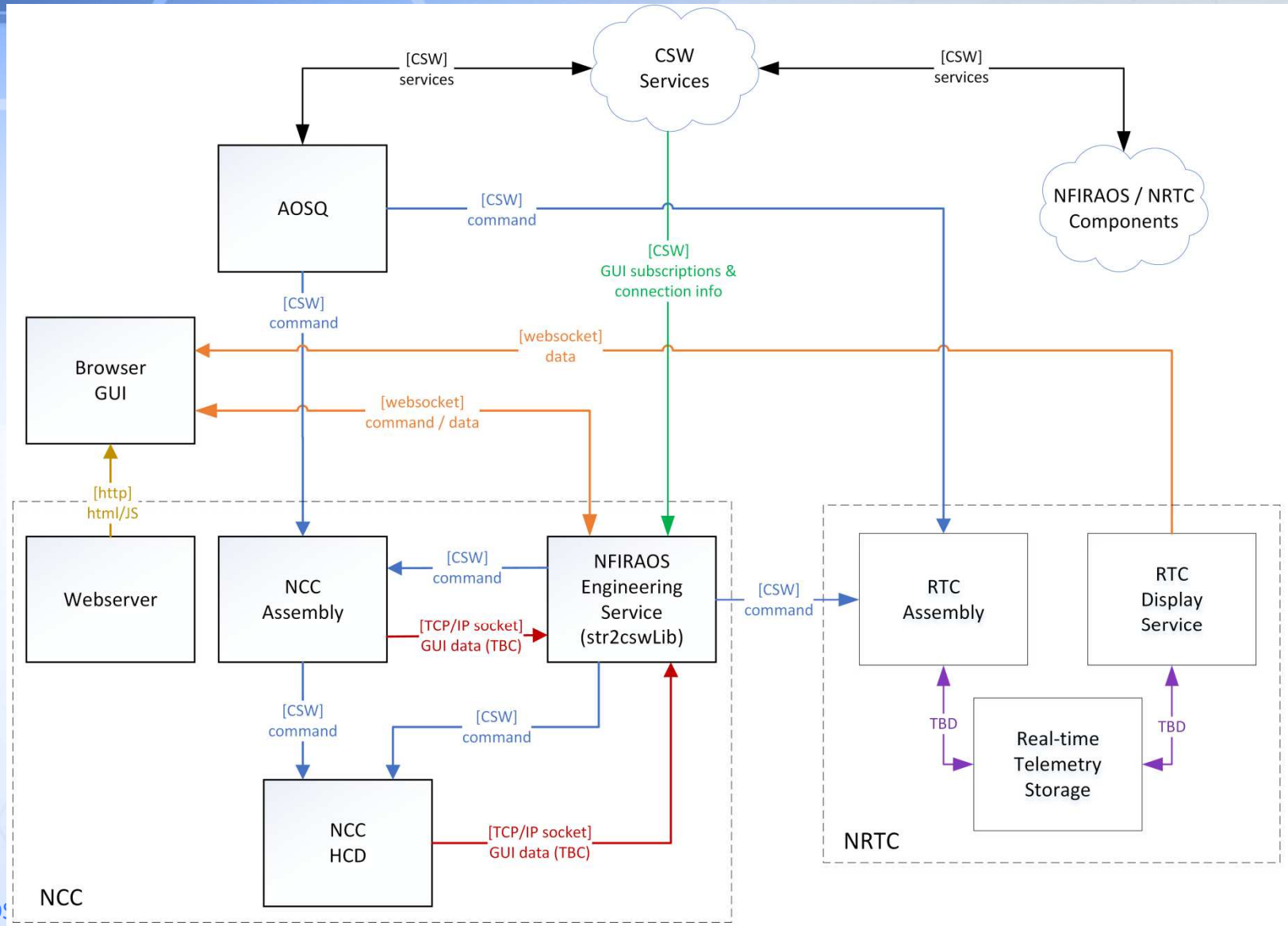
Real-time Telemetry Storage



TAC (Telemetry Analysis Computer)

- Used for diagnostic analysis of telemetry data
- Option of running TAC software on dual CPU Xeon server (faster single threaded performance) or KNL Xeon Phi (16 GB of high bandwidth memory)
  - If analysis only requires 16 GB RAM, then MCDRAM on Phi can be used by appropriate numactl command
  - 1 Phi could be configured to use 16 GB MCDRAM as cache
- It is possible to use more than one server to run more than one copy of TAC software

NCC & RTC Engineering GUIs



NFIRAOS CALSW Architecture