(What Would Have Been) Initial Experiences with a Cluster Mounted Flash File System

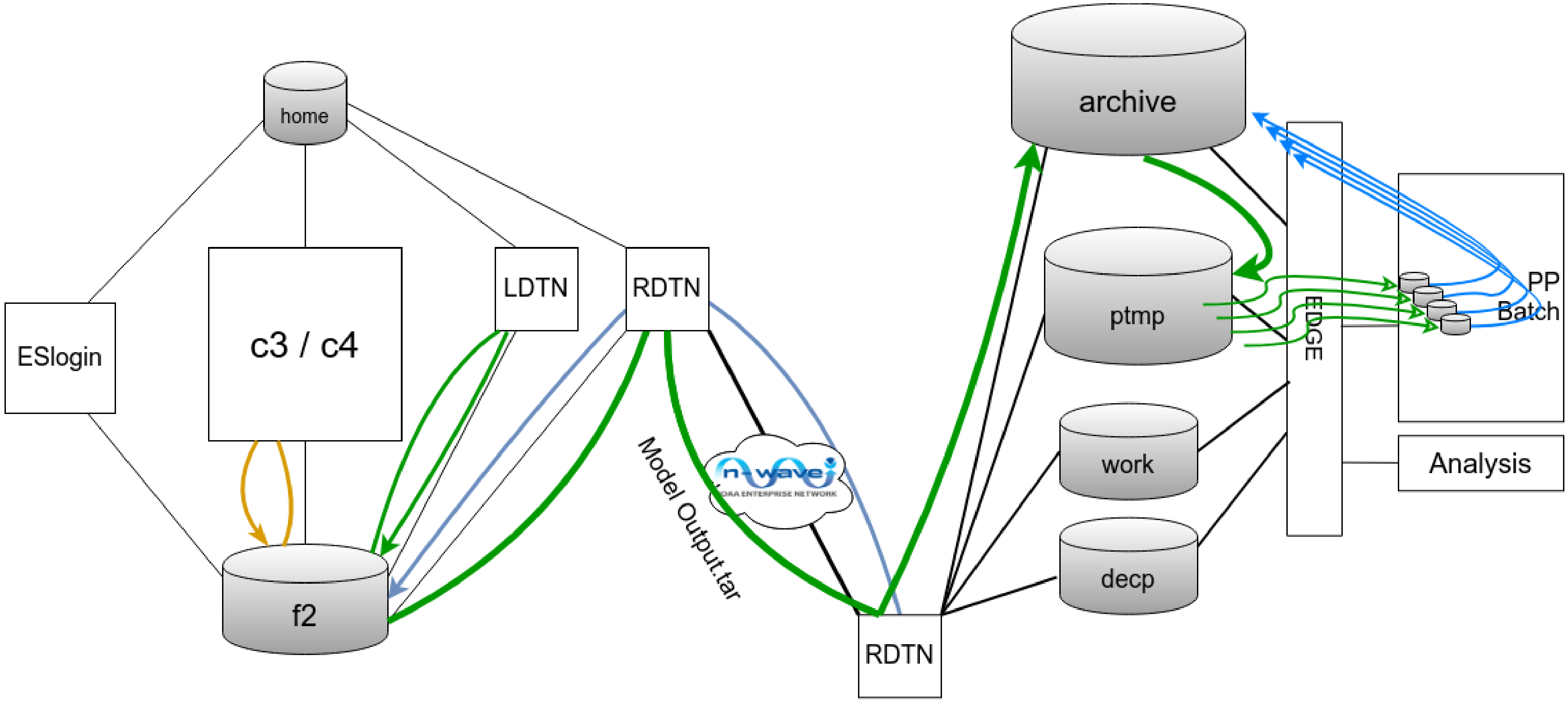
6th ENES HPC Workshop
(What should have been)
Hamburg, DE
29 May 2020

Jeff Durachta, PhD
Engineering Lead
Modeling Systems Division
Geophysical Fluid Dynamics Lab
Princeton, NJ

Overview

- Introduction
- Some Technology Fundamentals
  - Remote Direct Memory Access (RDMA)
  - Nonvolatile Memory express over Fabrics (NVMe-oF)
- Vast Data: A flash based storage appliance
- Towards Data Center Scale Computing
- How will we quantify, track & maintain “Better”?
  - EPMT: the Experiment Perf Metrics Tracking infrastructure
- Conclusions
- Acknowledgements

The Problem: Explosive Growth = Spiraling Complexity



- Increasing model resolution & complexity drives huge increases in data volumes
- Current workflows involve large amounts of data motion
- Workflows must adapt to deploy new concepts of storage and in-flight processing

Technology Fundamentals

- Remote Direct Memory Access (RDMA)
  - Read from and write to remote server memory with no involvement of CPUs, caches or system context switches
- RDMA is a standard protocol for InfiniBand networks
- RDMA over Converged Ethernet v2 (aka RoCE v2)
  - Provides routable packets (aka Routable RoCE)
  - Defines congestion control mechanism
  - Net Result: supports multi-domain networks
- Both InfiniBand and RoCE v2 have grown to support software architectures that consolidate compute, network and storage: i.e. the Hyperconverged Data Center

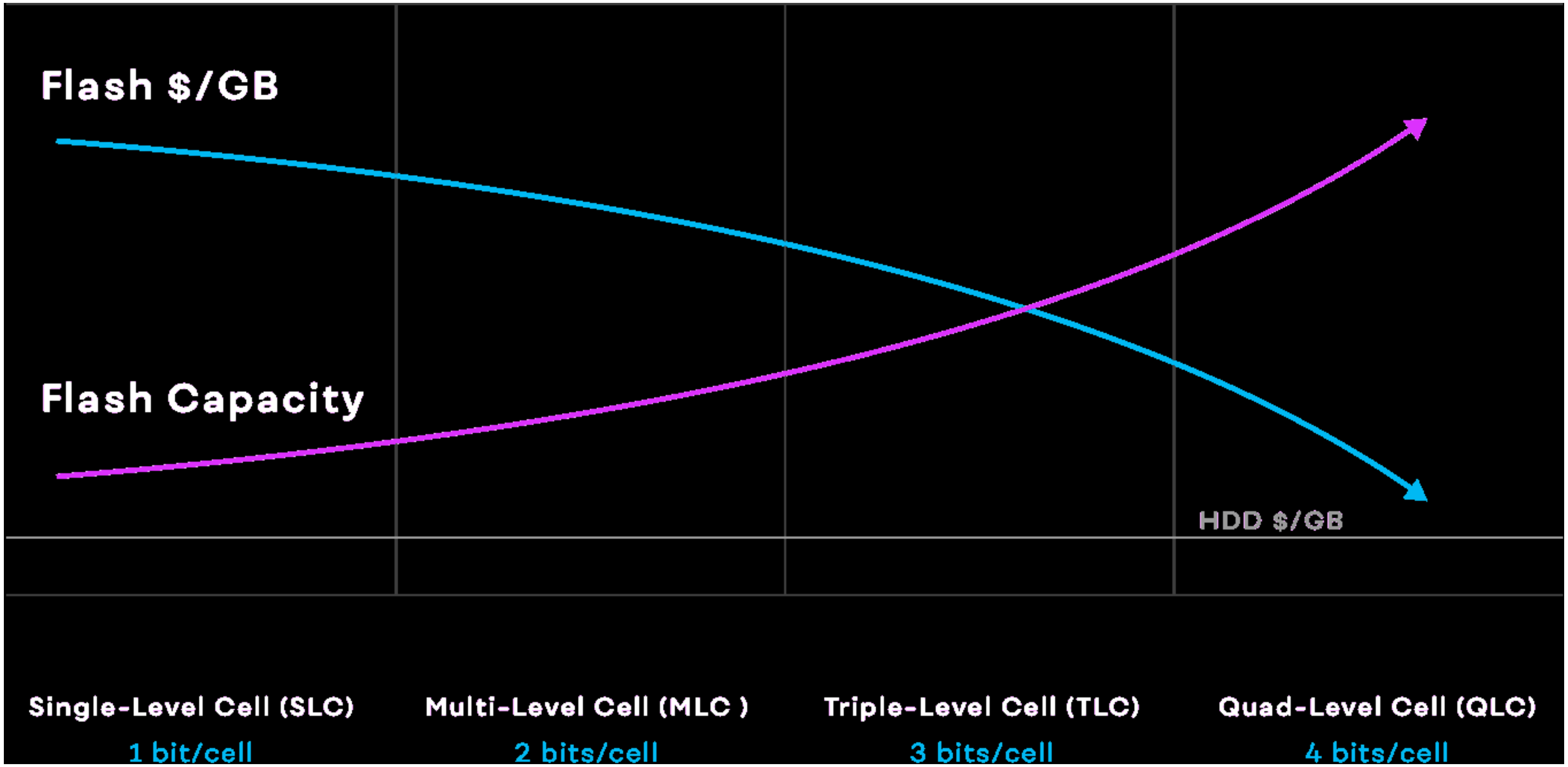
For an interesting discussion of InfiniBand and High Speed Ethernet, see: https://www.nextplatform.com/2019/02/19/ethernet-and-the-future-of-data-networking/

Non-Volatile Memory express (NVMe)

- NVMe: a S/W interface providing high levels of parallelism with lower queue overheads vs previous (SCSI) instruction set
  - Makes NVMe SSDs significantly faster than SAS or SATA components
- NVMe over Fabrics (NVMe-oF): extends NVMe protocols over commodity Ethernet and InfiniBand
- InfiniBand and Ethernet have ambitious performance hopes
  - **Each has a technology roadmap (dreamscape?) leading to the multi-Tbit/sec regime ~3 years**
- RDMA + NVMe-oF can bring access latency down to the microsecond regime
  - **RDMA + NVMe-oF enables Data Center / Hyperscale Computing**

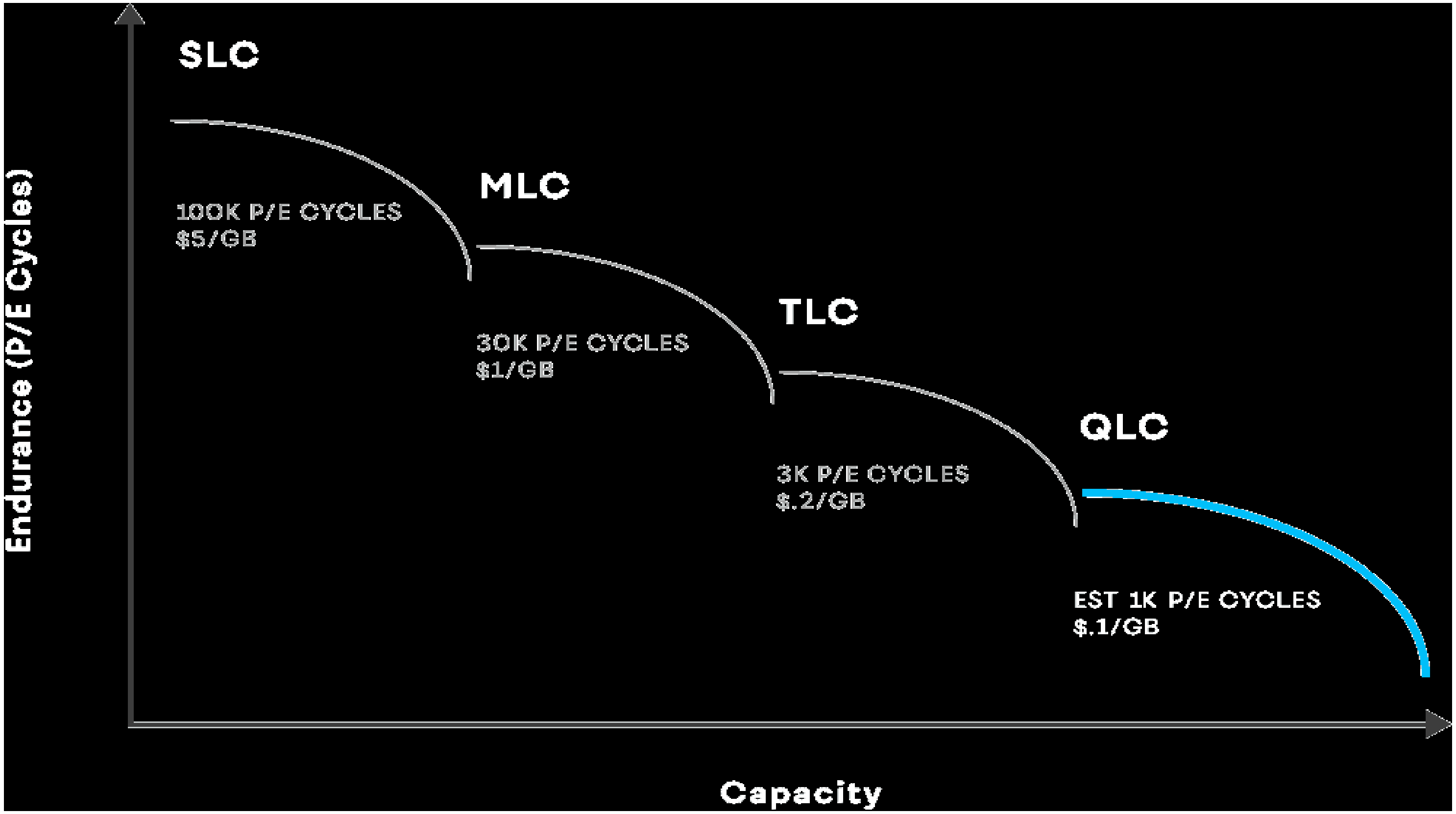
The VAST Data Storage Appliance

- Rides the economics of Quad-Level Cell Flash



Source: Vastdata.com

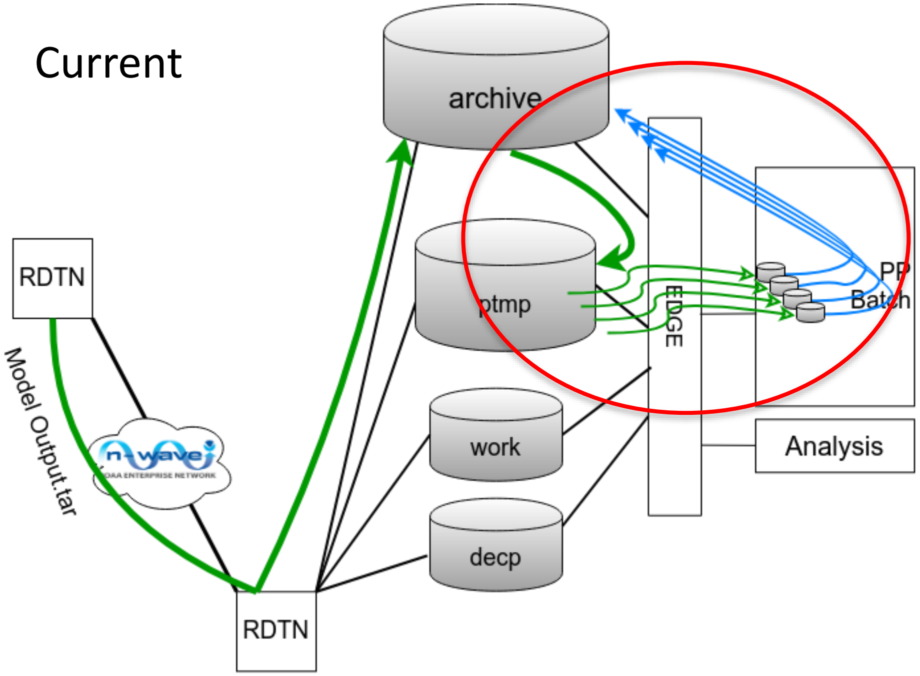
The VAST Data Storage Appliance



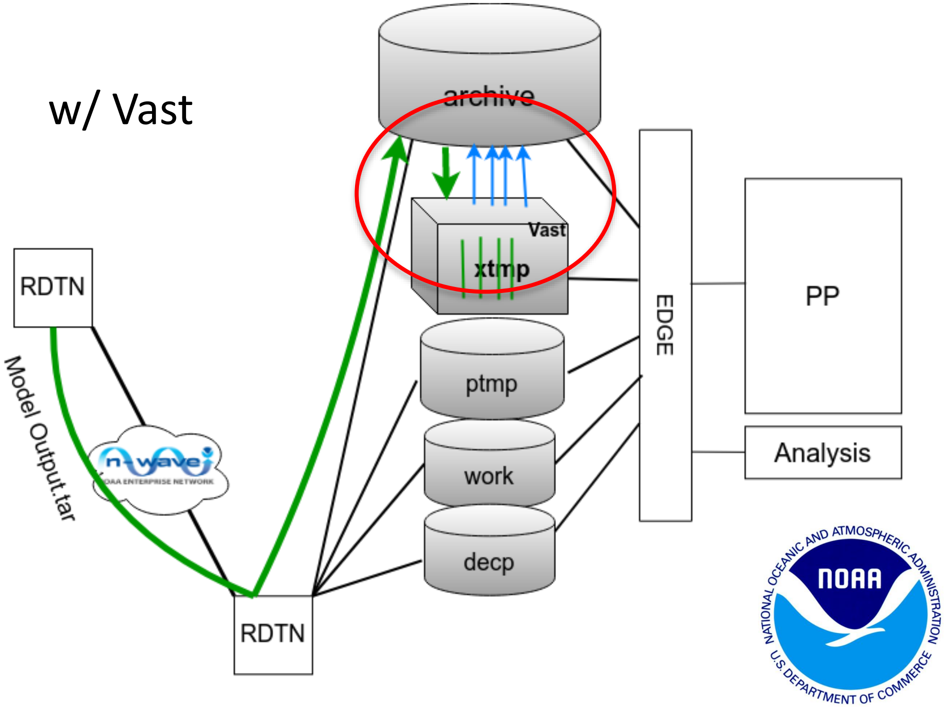
Source:
Vastdata.com

- Introduces unique technology to manage wear
  - Intel Optane 3D Xpoint memory provides global metadata store as well as a write buffer to minimize flash
  - The Vast file system compresses data to minimize use

The VAST @GFDL v1



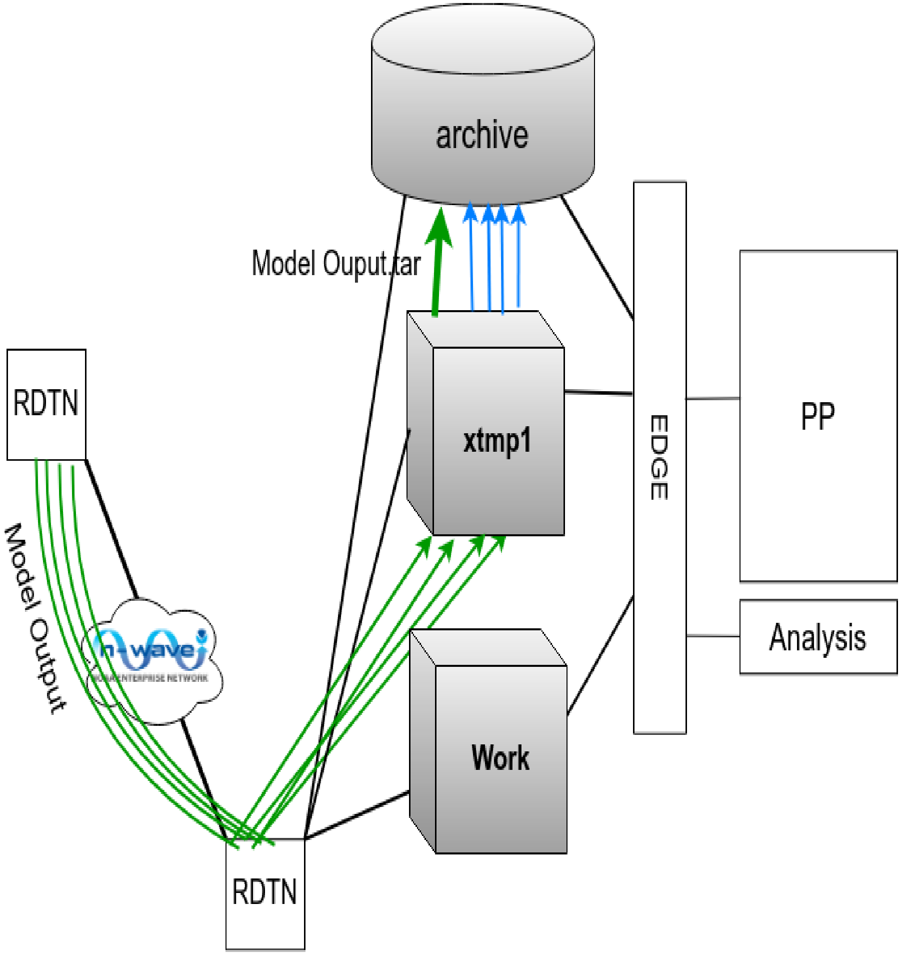
- The current workflow moves the same data to multiple stops
- Measurement shows 40-60% of post-processing wall clock is data movement



- Due to h/w limitations, 1st pass will continue to put diag tarball in archive then copy and untar to /xtmp (Vast)
- The data will then be post-processed in place and results put in archive

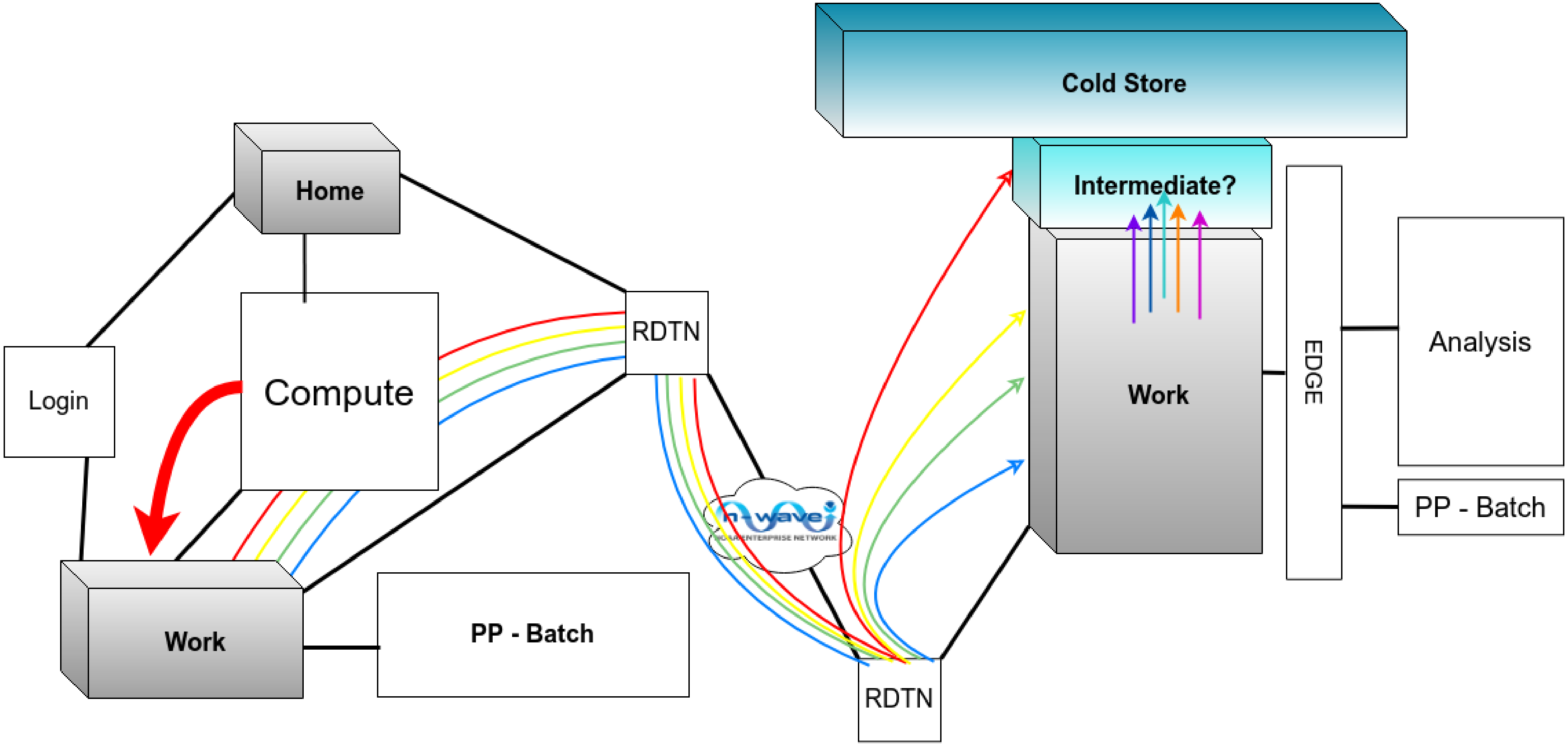
The VAST @GFDL v2?

- With proper hardware (e.g. local DTNs with RDMA), one might:
  - Send model component output files as a swarm to xtmp
  - Create and archive the raw model output tarball
  - At the same time, initiate postprocessing
- Does not reduce data motion of v1, but increases concurrency
- Does introduce potential B/W efficiency and workflow robustness issues



But of course, why stop here?

Towards In-Flight Post-Processing



- Ubiquitous availability of NVMe-oF technology will enable new storage hierarchies and data in-flight
- In turn, this will enable Earth System Modeling workflows to achieve “Data Center Scale Computing”
- Potential for B/W efficiency and workflow robustness issues remain

Quantify, Track & Maintain “Better”

- While future technologies may improve many things
  - More things “in-flight” ⇔ more opportunities for failures & performance degradation
- 2018: 5th ENES HPC Workshop, Lecce, IT
  - Towards HPC System Throughput Optimization
  - “Workflow Data is Everywhere”
  - Initial thoughts on WorkflowDB model
- As of 30 June, 2020
  - Final stages of deploying data gathering and performance analysis infrastructure v1.0 on GFDL PP and Analysis:

EPMT: Experiment Performance Metrics Tracking

Selected EPMT v1.0 Capabilities

- Fully transparent to the user and portable to any h/w platform running current versions of Linux
- Developed from many open source projects
- User definable, DB searchable job tags
- Automated data ingestion, aging, retirement, migration
- Complete API for exploration and analysis of data
- IPython notebooks documenting APIs and performance analyses of post-processing experiments
- Advanced classification, outlier detection & dimensionality reduction techniques
  - Single & multi-variate outlier detection w/ ref model development
  - Root Cause Analysis tools and techniques
  - Principle Component Analysis for feature importance and reduction
- GUI with drill down interactive visualizations

Conclusions

- Increasing model resolution & complexity are driving huge increases in data volumes
  - Workflows must adapt to deploy new concepts of storage and in-flight processing
- RDMA + NVMe-oF provides fundamental technology supporting Data Center / Hyperscale Computing
- Properly handled, Flash NAND can provide revolutionary improvements and economies in the storage hierarchy
- Putting this all together, we see a path forward for Data Center Scale Computing for Earth System Modeling
- These new levels of sophistication bring additional complexity & greater need to know what's driving workflow performance
  - The Experiment Perf Metrics Tracking (EPMT) infrastructure provides a way forward for workflow performance data and analytics

Acknowledgements

The GFDL Workflow Team:

V Balaji[1], Chris Blanton[2], Jeff Durachta[3], Tom Jackson[2], Sergey Nikonov[1], Aparna Radhakrishnan[1], Kris Rand[2], Luis Sal-Bey[2], Seth Underwood[3], Hans Vahlenkamp[4], Chan Wilson[2]

EPMT Development[5]

Philip Mucci, Tushar Mohan, Chris Ault

[1] *Princeton University*

[2] *SAIC*

[3] *US Federal Government*

[4] *University Corporation for Atmospheric Research (UCAR)*

[5] *Minimal Metrics, LLC*