12th ANNUAL WORKSHOP 2016

HPC STORAGE AND IO TRENDS AND WORKFLOWS

Gary Grider, Division Leader, HPC Division

Los Alamos National Laboratory

April 4, 2014

LA-UR-16-20184

EIGHT DECADES OF PRODUCTION WEAPONS COMPUTING TO KEEP THE NATION SAFE

Maniac

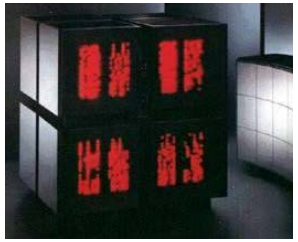
IBM Stretch

CDC

Cray 1

Cray X/Y

CM-2



CM-5

SGI Blue Mountain

DEC/HP Q

IBM Cell Roadrunner

Cray XE Cielo

Cray Intel KNL Trinity

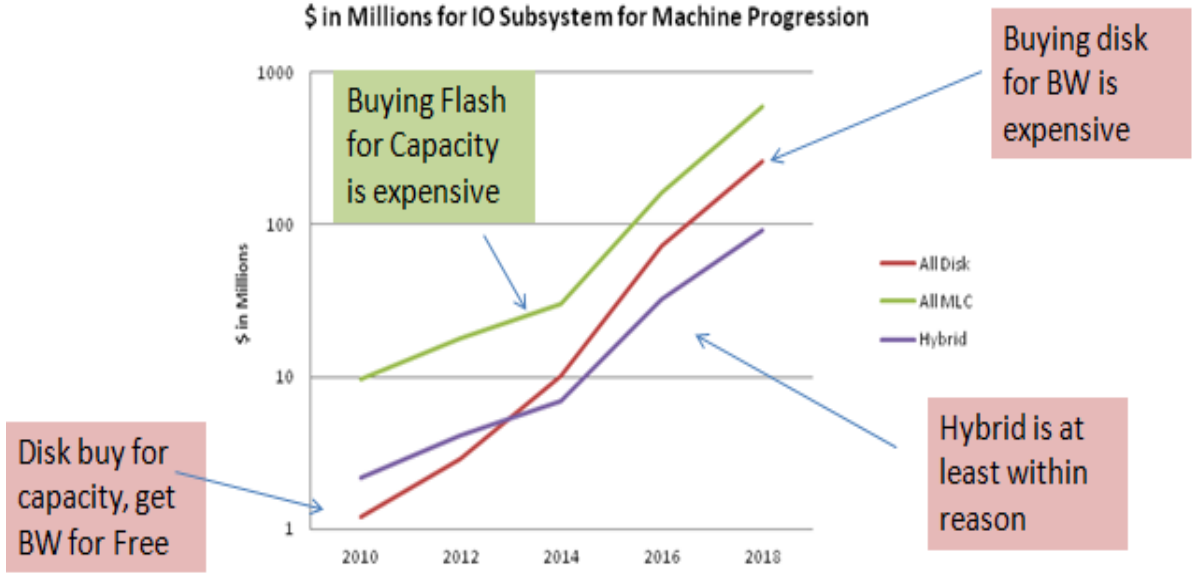
Ziggy DWave

Cross Roads

CROSS ROADS
An APEX Collaboration

ECONOMICS HAVE SHAPED OUR WORLD

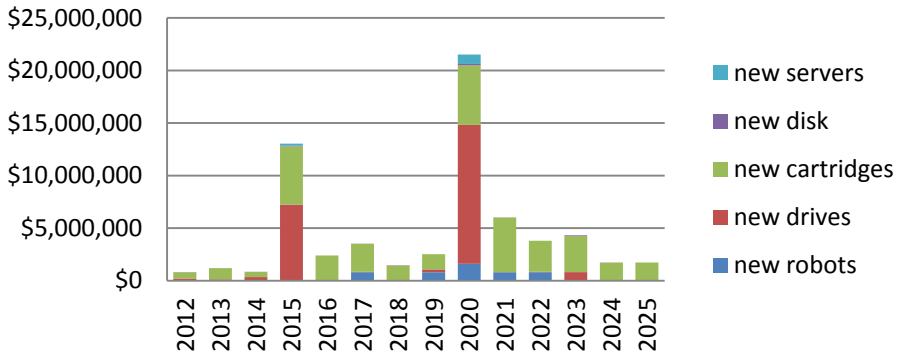
The beginning of storage layer proliferation circa 2009



- **Economic modeling for large burst of data from memory shows bandwidth / capacity better matched for solid state storage near the compute nodes**

- **Economic modeling for archive shows bandwidth / capacity better matched for disk**

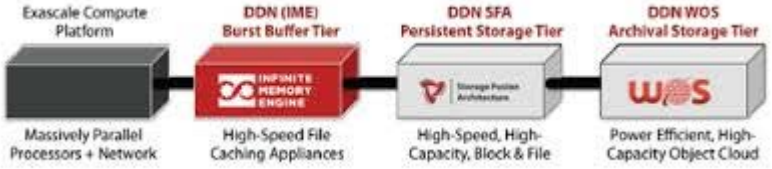
Hdwr/media cost 3 mem/mo 10% FS



THE HOOPLA PARADE CIRCA 2014

DDN STORAGE

INFINITE MEMORY ENGINE™

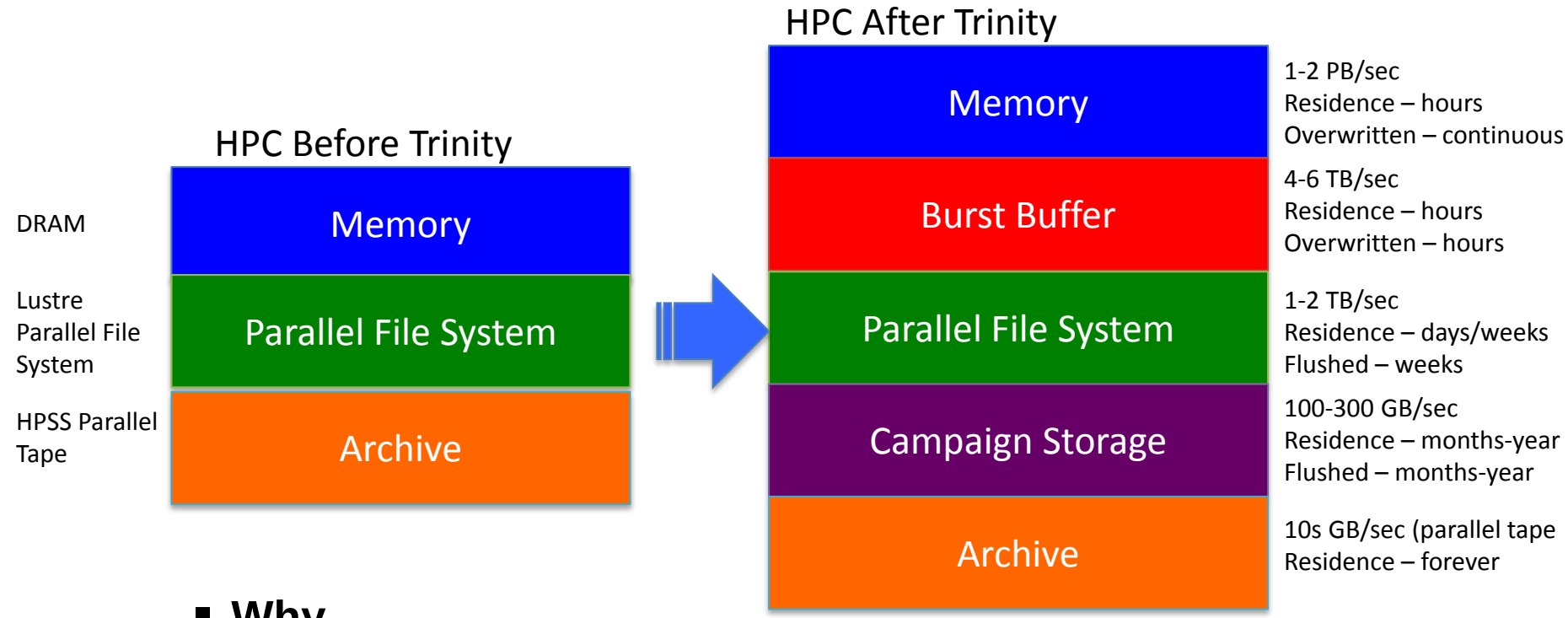


CRAY
THE SUPERCOMPUTER COMPANY

Data Warp

Powered By Dilithium Crystals

WHAT ARE ALL THESE STORAGE LAYERS? WHY DO WE NEED ALL THESE STORAGE LAYERS?



- **Why**
  - BB: Economics (disk bw/iops too expensive)
  - PFS: Maturity and BB capacity too small
  - Campaign: Economics (tape bw too expensive)
  - Archive: Maturity and we really do need a “forever”

BURST BUFFERS ARE AMONG US

- **Many instances in the world now, some at multi-PB Multi-TB/s scale**
- **Uses**
  - Checkpoint/Restart, Analysis, Out of core
- **Access**
  - Largely POSIX like, with JCL for stage in, stage out based on job exit health, etc.
- **A little about Out of Core**
  - Before HBM we were thinking DDR ~50-100 GB/sec and NVM 2-10 GB/sec (10X penalty and durability penalty)
    - Only 10X penalty from working set speed to out of core speed
  - After HBM we have HBM 500-1000 GB/sec, DDR 50-100 GB/sec, and NVS 2-10 GB/sec (100X penalty from HBM to NVS)
    - Before HBM, out of core seemed like it might help some read mostly apps
    - After HBM, using DDR for working set seems limiting, but useful for some
    - Using NVS for read mostly out of core seems limiting too, but useful for some

CAMPAIGN STORAGE SYSTEMS ARE AMONG US TOO – MARFS MASSIVELY SCALABLE POSIX LIKE FILE SYSTEM NAME PACE OVER CLOUD STYLE ERASURE PROTECTED OBJECTS

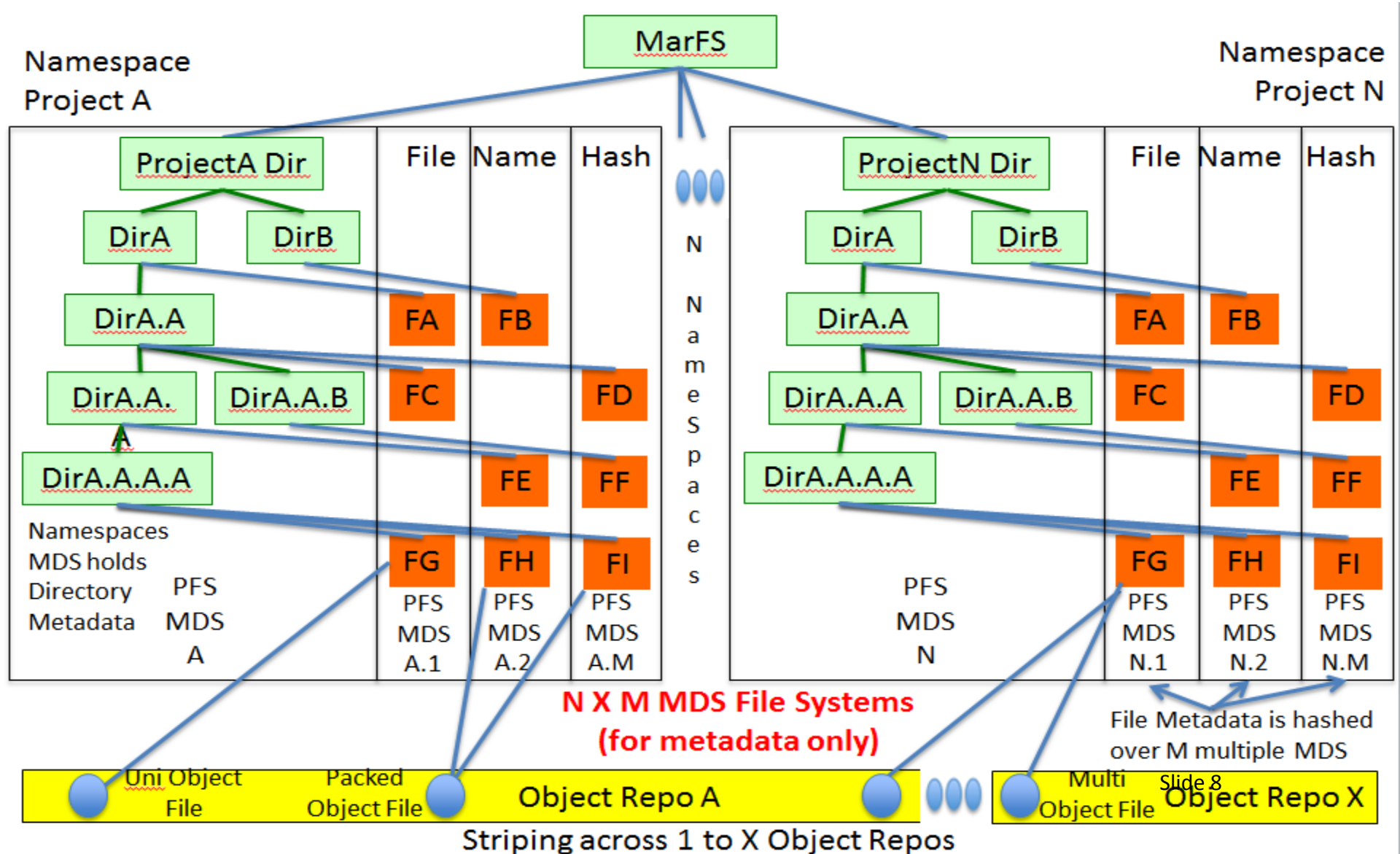
- **Background**
  - Object Systems provide massive scaling and efficient erasure
  - Friendly to applications, not to people. People need a name space.
  - Huge Economic appeal (erasure enables use of inexpensive storage)
  - POSIX name space is powerful but has issues scaling
- **The challenges**
  - Mismatch of POSIX an Object metadata, security, read/write size/semantics
  - No update in place with Objects
  - Scale to Trillions of files/directories and Billions of files in a directory
  - 100's of GB/sec but with years data longevity
- **Looked at**
  - GPFS, Lustre, Panasas, OrangeFS, Cleversafe/Scality/EMC ViPR/Ceph/Swift, Glusterfs, Nirvana/Storage Resource Broker/IRODS, Maginatics, Camlistore, Bridgestore, Avere, HDFS
- **Experiences**
  - Pilot scaled to 3PB and 3 GB/sec,
  - First real deployment scaling to 30PB and 30 GB/sec
- **Next a demo of scale to trillions and billions**

Current Deployment Uses N GPFS's for MDS and Scality Software Only Erasure Object Store

Be nice to the Object system: pack many small files into one object, break up huge files into multiple objects

Open Source, BSD License
Partners Welcome
https://github.com/mar-file-system/marfs
https://github.com/pftool/pftool)

MARFS
METADATA SCALING NXM
DATA SCALING BY X



ISN'T THAT TOO MANY LAYERS JUST FOR STORAGE?

RIP PFS ?

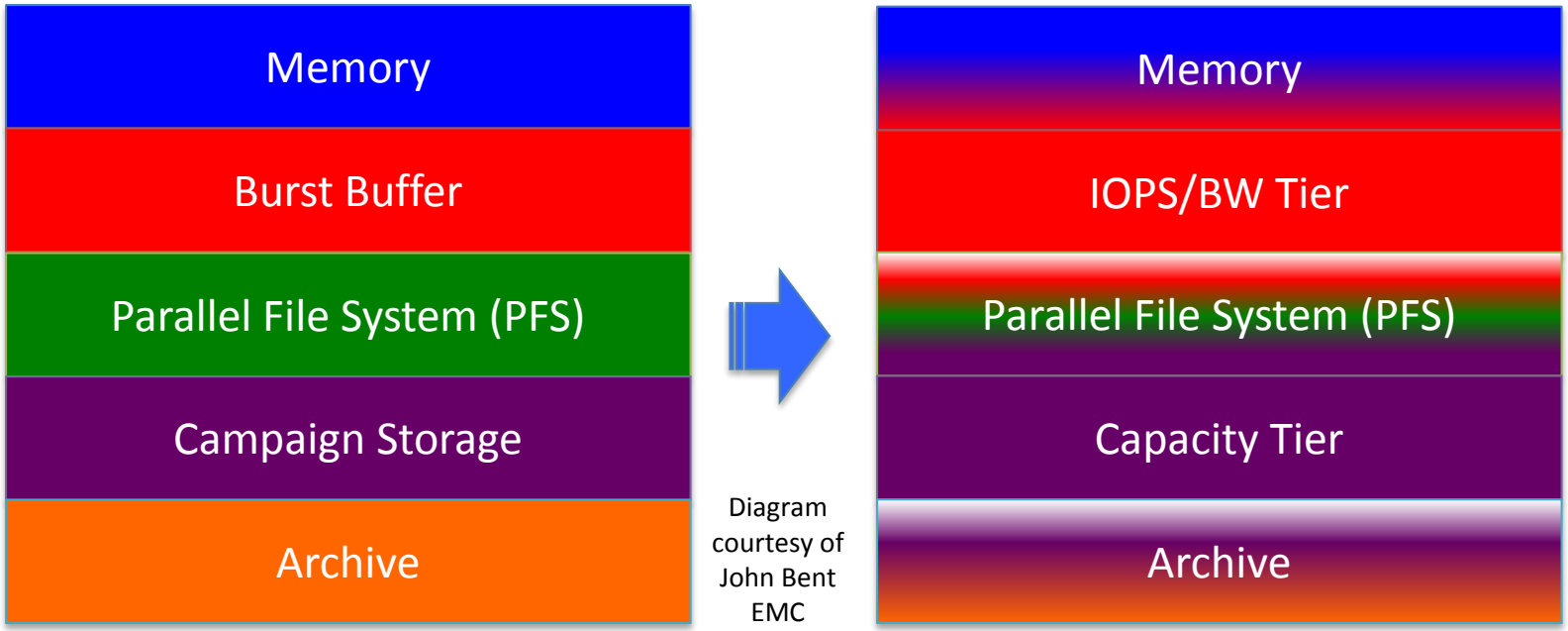


Diagram courtesy of John Bent EMC

Factoids (times are changing!)

LANL HPSS = 53 PB and 543 M files

Trinity 2 PB memory, 4 PB flash (11% of HPSS) and 80 PB PFS or 150% HPSS)

Crossroads may have 5-10 PB memory, 40 PB solid state or 100% of HPSS with data residency measured in days or weeks

- **If the Burst Buffer does its job very well (and indications are capacity of in system NV will grow radically) and campaign storage works out well (leveraging cloud), do we need a parallel file system anymore, or an archive? Maybe just a bw/iops tier and a capacity tier.**
- **Too soon to say, seems feasible longer term**

We would have never contemplated more in system storage than our archive a few years ago

BURST BUFFER -> PERF/IOPS TIER WHAT WOULD NEED TO CHANGE?

- **Burst Buffers are designed for data durations in hours-days. If in system solid state in system storage is to be used for months duration many things are missing.**
  - Protecting Burst Buffers with RAID/Erasure is NOT economical for checkpoint and short term use because you can always go back to a lower tier copy, but longer duration data requires protection. Likely you would need a software RAID/Erasure that is distributed on the Supercomputer over its fabric
  - Long term protection of the name space is also needed
  - QoS issues are also more acute
  - With much more longer term data, locality based scheduling is also perhaps more important

CAMPAIGN -> CAPACITY TIER
WHAT WOULD NEED TO CHANGE CAMPAIGN?

- **Campaign data duration is targeted at a few years but for a true capacity tier more flexible perhaps much longer time periods may be required.**
  - Probably need power managed storage devices that match the parallel BW needed to move around PB sized data sets
- **Scaling out to Exabytes, Trillions of files, etc.**
  - Much of this work is underway
- **Maturing of the solution space with multiple at least partially vendor supported solutions is needed as well.**

OTHER CONSIDERATIONS

- **New interfaces to storage that preserve/leverage structure/semantics of data (much of this is being/was explored in the DOE Storage FFwd)**
  - DAOS like concepts with name spaces friendly to science application needs
  - Async/transactional/Versioning to match better with future async programming models
- **The concept of a loadable storage stack (being worked in MarFS and EMC)**
  - It would be nice if the Perf/IOPS tier could be directed to “check out” a “problem” to work on for days/weeks. Think PBs of data and billions of metadata entries of various shapes. (EMC calls this dynamically loadable name spaces)
  - MarFS metadata demonstration of Trillions of files in a file system and Billions of files in a directory will be an example of a “loadable” name space
  - CMU’s IndexFS->BatchFS->DeltaFS - stores POSIX metadata into thousands of distributed KVS’s makes moving/restart with billions of metadata entries simply and easily
  - Check out a billion metadata entries, make modifications, check back in as a new version or a merge back into the Capacity Tier master name space

BUT, THAT IS JUST ECONOMIC ARM WAVING.

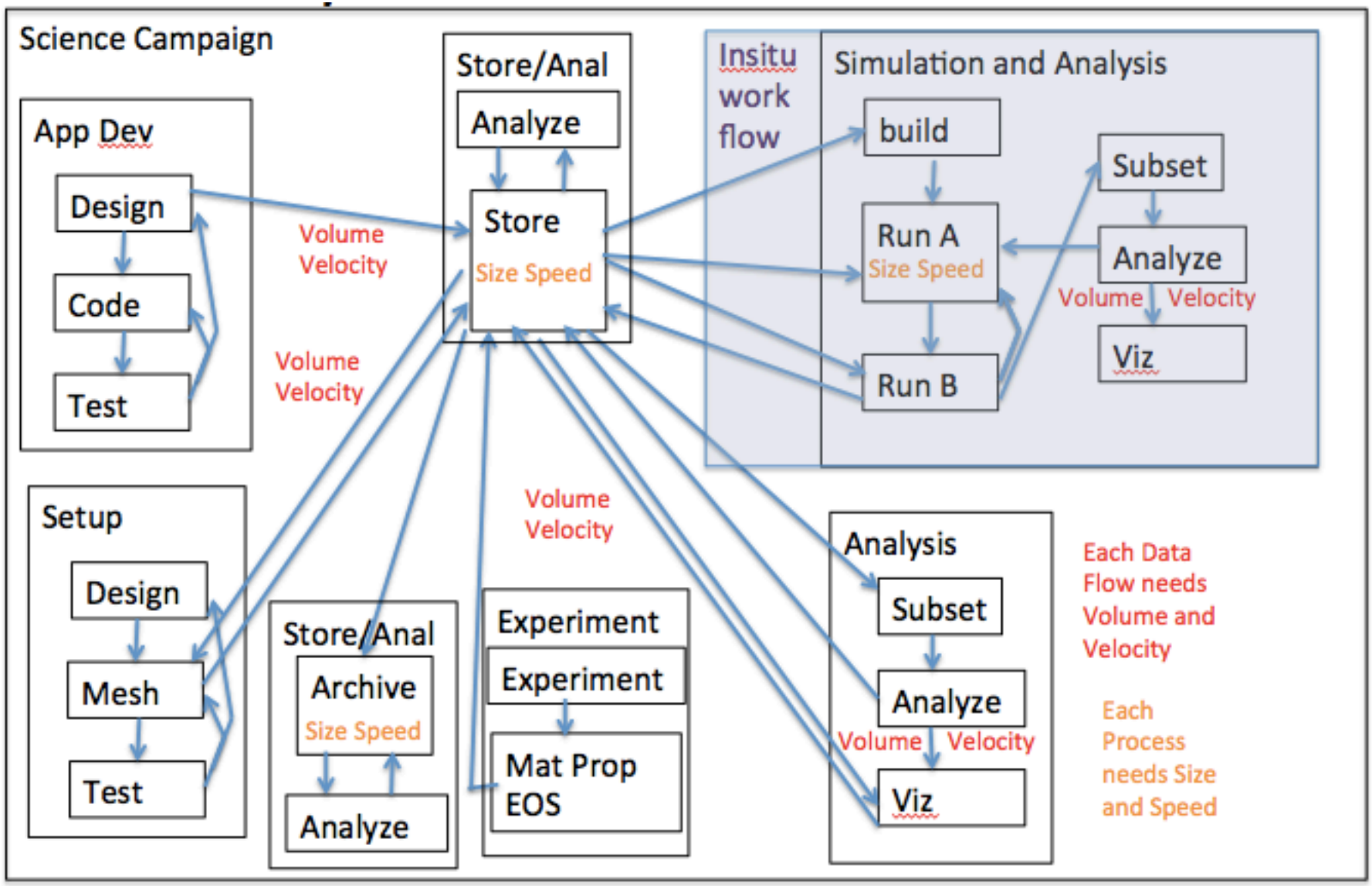
How will the economics combine with the apps/machine/environmental needs?

Enter Workflows

WORKFLOWS TO THE RESCUE?

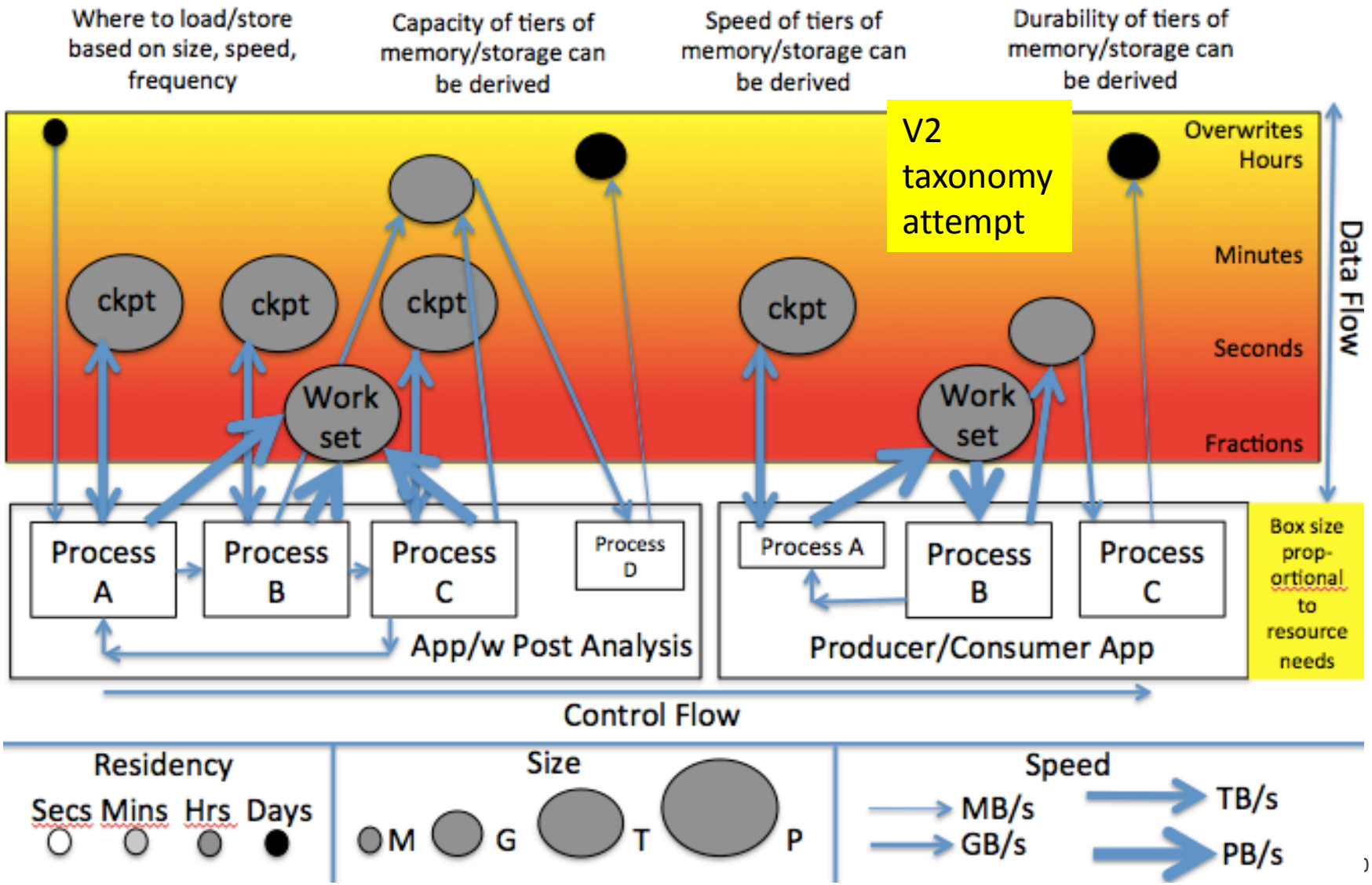
- **What did I learn from the workflow-fest circa 04/2015?**
  - There are 57 ways to interpret in situ ☺
  - There are more workflow tools than requirements documents
  - There is no common taxonomy that can be used to reason about data flow/workflow for architects or programmers ☹
- **What did I learn from FY15 Apex Vendor meetings**
  - Where do you want your flash, how big, how fast, how durable
  - Where do you want your SCM, how big, how fast
  - Do you want it near the node or near the disk or in the network
  - --- YOU REALLY DON'T WANT ME TO TELL YOU WHERE TO PUT YOUR NAND/SCM ---
- **Can workflows help us beyond some automation tools?**

INSITU / POST / ACTIVE STORAGE / ACTIVE ARCHIVE ANALYSIS WORK FLOWS IN WORK FLOWS



WORKFLOWS: POTENTIAL TAXONOMY

Derived from Dave Montoya (Circa 05/15)



WORKFLOWS CAN HELP US BEYOND SOME AUTOMATION TOOLS: WORKFLOWS ENTER THE REALM OF RFP/PROCUREMENT

- **Trinity/Cori**
  - We didn't specify flops, we specified running bigger app faster
  - We wanted it to go forward 90% of the time
  - We didn't specify how much burst buffer, or speeds/feeds
  - Vendors didn't like this at first but realized it was degrees of freedom we were giving them
- **Apex Crossroads/NERSC+1**
  - Still no flops ☺
  - Still want it to go forward a large % of the time
  - Vendors ask: where and how much flash/nvram/pixy dust do we put on node, in network, in ionode, near storage, blah blah
  - We don't care we want to get the most of these work flows through the system in 6 months

V3 Taxonomy Attempt

Next slides represents work done by the APEX Workflow Team

V3 TAXONOMY FROM APEX PROCUREMENT DOCS
A SIMULATION PIPELINE

APEX Workflows

LANL, NERSC, SNL

SAND2015-10342 O
LA-UR-15-29113

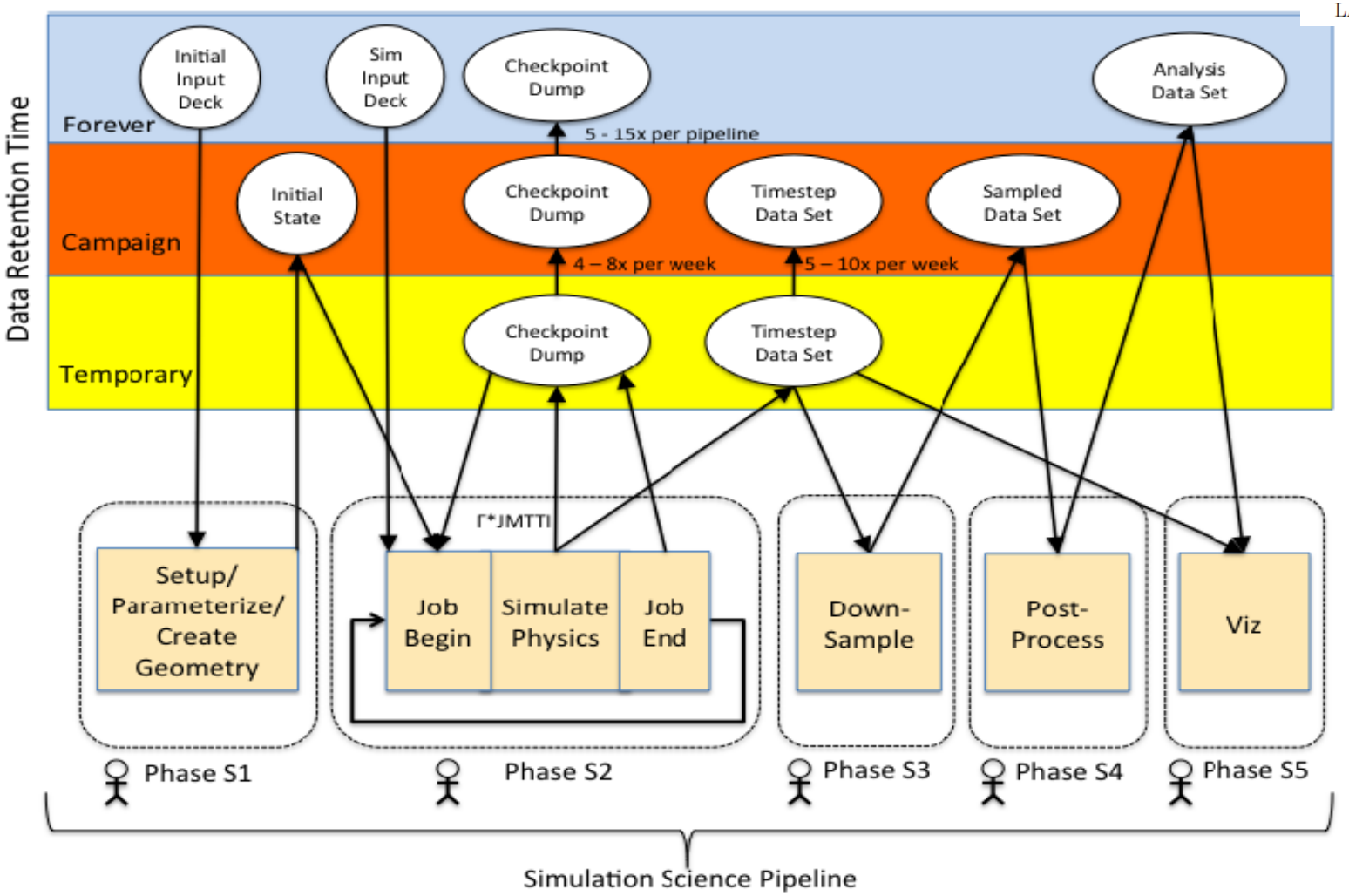


Figure 1: An example of an APEX simulation science workflow.

V3 TAXONOMY FROM APEX PROCUREMENT DOCS A HIGH THROUGHPUT/UQ PIPELINE

APEX Workflows

LANL, NERSC, SNL

SAND2015-10342 O
LA-UR-15-29113

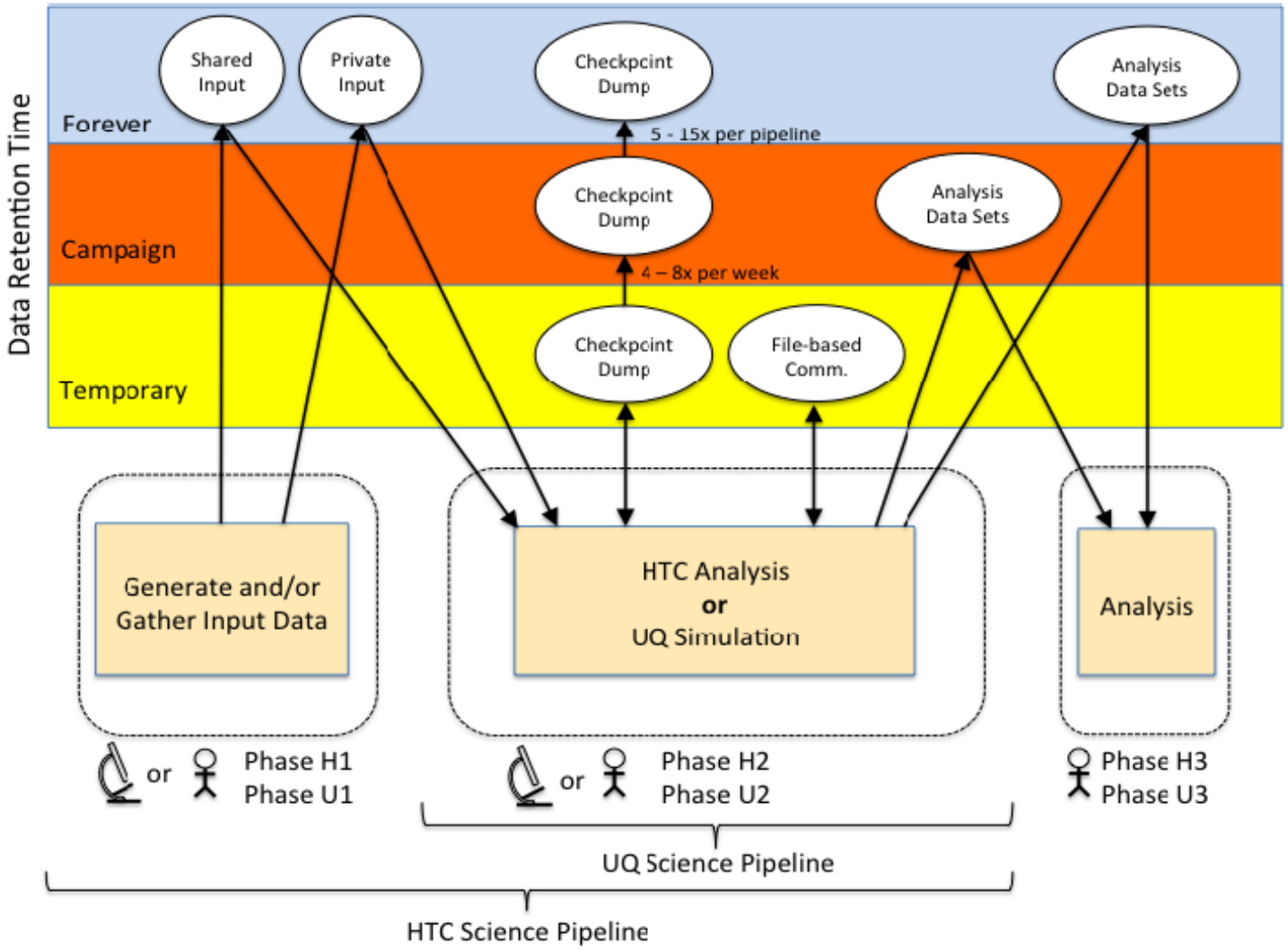


Figure 2: Example HTC and UQ workflow

WORKFLOW DATA THAT GOES WITH THE WORKFLOW DIAGRAMS

	Tri-Labs workload							NERSC workload						
	LANL				SNL		LLNL							
Workflow	EAP	LAP	Silverton	VPIC	Dakota A	Dakota S	Pending	ALS	CESM	GTS	HipMer	Materials	MILC	Sky Survey
Workflow type	Sim	Sim	Sim	Sim	Sim/UQ	UQ	Pending	HTC	Sim	Sim	HTC	Sim	Sim	HTC
Site Workload percentage	60	5	15	10	10 to 15	10 to 15	100	< 1	3	1	< 1	7	5	< 1
Representative workload percentage	20	2	5	3	4	4	33	3	6	6	7	19	11	3
Number of Cielo cores	65536	32768	131072	70000	131072	65536		100	2064	16384	960	2400	100000	24
Number of workflow pipelines per allocation	1 to 15	1 to 5	1 to 10	5 to 10	100 to 1000	50 to 200		10760	8	100	100	100	1000	21000
Number of simultaneous allocations	20 to 25	2 to 3	2 to 3	2 to 3	5 to 10	5 to 10	8							
Anticipated increase in problem size by 2020	10 to 12x	8 to 12x	6x	10x	2 to 4x	1x		1x	16 to 23x	5x	1x	10 to 25x	1x	1x
Anticipated increase in workflow pipelines per allocation by 2020	1x	1x	1x	1x	2 to 8x	2x		5x	3x	1x	50x	1x	?	2.38x
Data retained (percentage of memory)	910.00	3050.00	1205.00	545.25	415.00	25.07		285.63	835.37	15.57	100.54	135.42	103.38	11.57
During pipeline	**50.00**	**85.00**	**320.00**	**222.75**	**20.00**	**0.15**		**147.68**	**0.29**	**0.68**	**34.34**	**20.83**	**102.53**	**2.16**
Analysis	20.00	10.00	5.00	200.00	5.00			126.57			34.34	20.83		2.16
Checkpoint	30.00	75.00	210.00	18.75	10.00	0.15			0.29	0.68				
Input			70.00	5.00	5.00	0.00		21.10						
Out-of-core													102.53	
During Allocation	**240.00**	**210.00**	**480.00**	**142.50**	**345.00**	**0.00**		**21.10**					**0.62**	
Analysis	60.00	60.00	60.00	100.00	40.00			21.10					0.62	
Checkpoint	180.00	150.00	420.00	37.50	300.00									
Input	0.00	0.00		5.00	5.00	0.00								
Forever	**620.00**	**2755.00**	**405.00**	**180.00**	**40.00**	**24.93**		**116.84**	**835.08**	**14.89**	**66.20**	**114.58**	**0.23**	**9.41**
Analysis	500.00	2500.00	5.00	130.00	35.00	24.44		106.29	808.14	14.89	0.36	114.58	0.12	0.62
Checkpoint	100.00	250.00	400.00	50.00		0.49			26.94					
Input	20.00	5.00			5.00			10.55	0.00		65.83	0.00	0.12	8.79

SUMMARY

- **Economic modeling/analysis is a powerful tool for guiding our next steps**
- **Given the growing footprint of Data Mgmt/Movement in the cost and pain in HPC, workflows may grow in importance and may be more useful in planning for new machine architectures/procurement/integration than ever.**
- **Combining Economics and Workflows helps paint a picture of the future for all of us.**

OPENFABRICS ALLIANCE

12th ANNUAL WORKSHOP 2016

THANK YOU
AND
RIP PFS