LANL Computing, From Architecture Guided Apps to Apps Guided Architectures 2021

Delivering science and technology
to protect our nation
and promote world stability

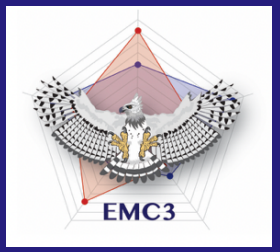
LA-UR-21-22315

Transforming Weapons Performance Calculation via Efficiency Mission Computing

- Responsible for care for most of the US Nuclear Stockpile
- A Serious Enduring Mission
- Not a solved problem space or anywhere close

Background

Eight Decades of Production Weapons Computing to Keep the Nation Safe



Maniac

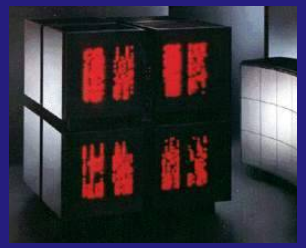
IBM Stretch

CDC

Cray 1

Cray X/Y

CM-2



CM-5

SGI Blue Mountain

DEC/HP Q

IBM Cell Roadrunner

Cray XE Cielo

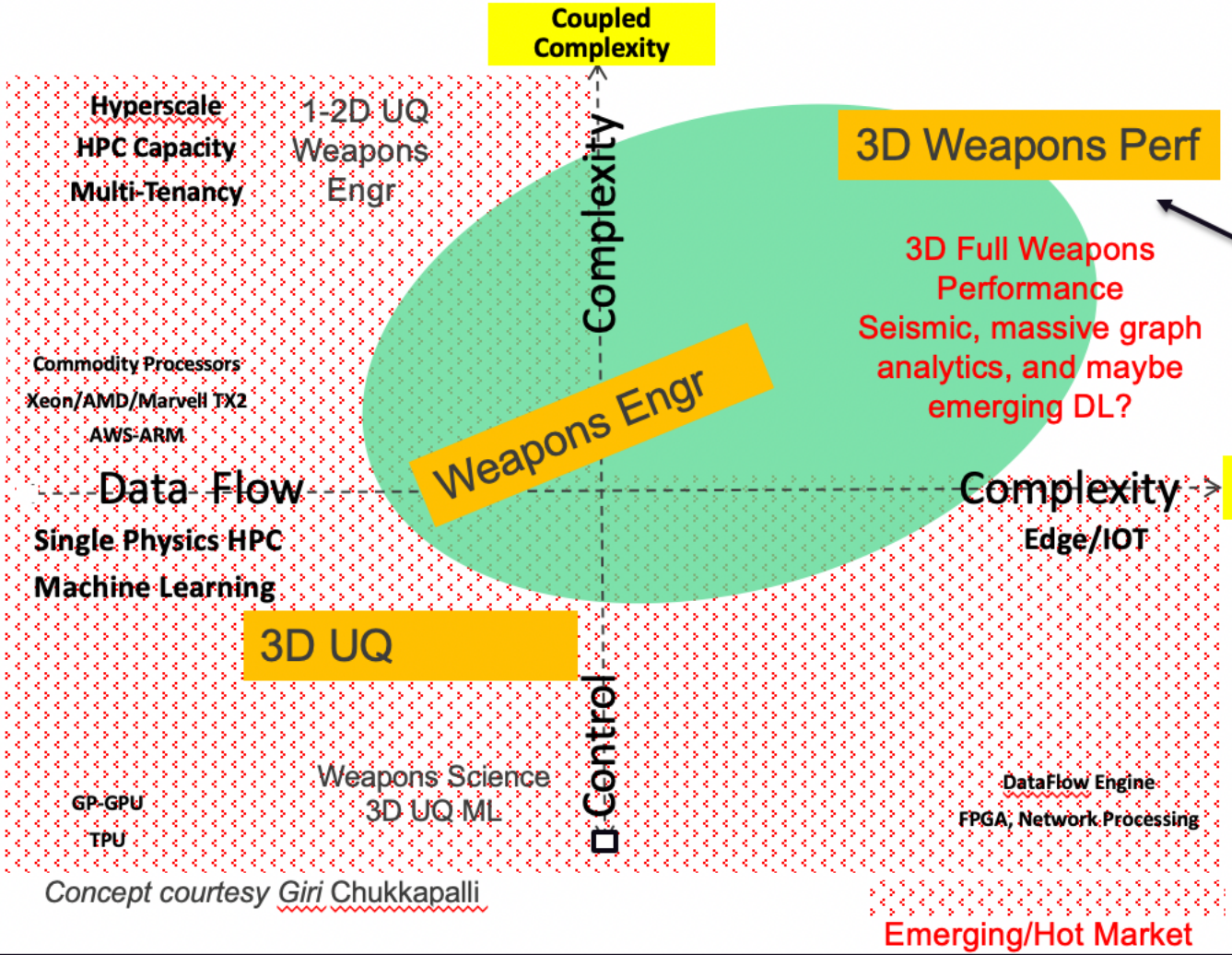
Cray Intel KNL Trinity

Ising DWave

Cross Roads

CROSS ROADS
An APEX Collaboration

Different Labs have different Mission Focus



All labs work in all 4 quadrants but LANL mission focus is here

How many sites dedicate ½ of a premier machine to a single problem for > 7months for a result (PB dram/10k nodes for most of a year!)

3D UQ (uncertainty Quantification) – many small simpler 3D weapons sims to generate statistically significant variance analysis - 1000's of TB sized 3D runs of size and complexity (each - roughly state of the art in 2004)

Current Site Info

Trinity – the current NNSA work horse

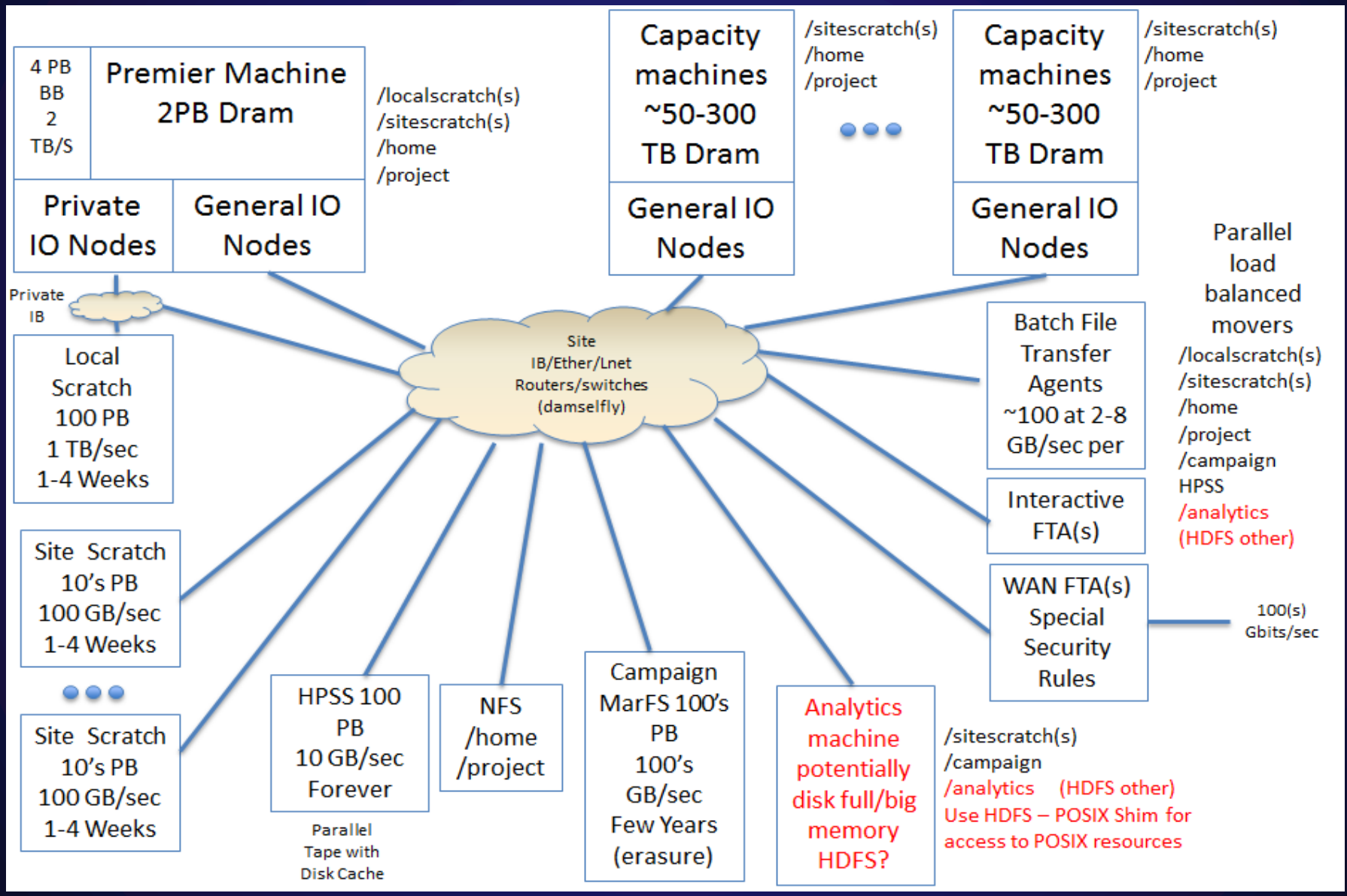
Trinity and Burst Buffer credited for enabling solution to decades old weapons issue that would not run on Cielo or Sequoia (1/4-1/2 machine for 7 months!)

- **Haswell and KNL**
- **20,000 Nodes**
- **Million-ish cores**
- **2 PByte DRAM**
- **4 PByte NAND Burst Buffer ~ 4 Tbyte/sec**
- **100 Pbyte Scratch CMR Disk File system ~1.2 Tbyte/sec**
- **60PB Sitewide SMR Disk Campaign Store (50 Gbyte/sec currently, Tiered Erasure Protected)**
- **80 PByte Sitewide Parallel Tape Archive ~ 3 Gbyte/sec**

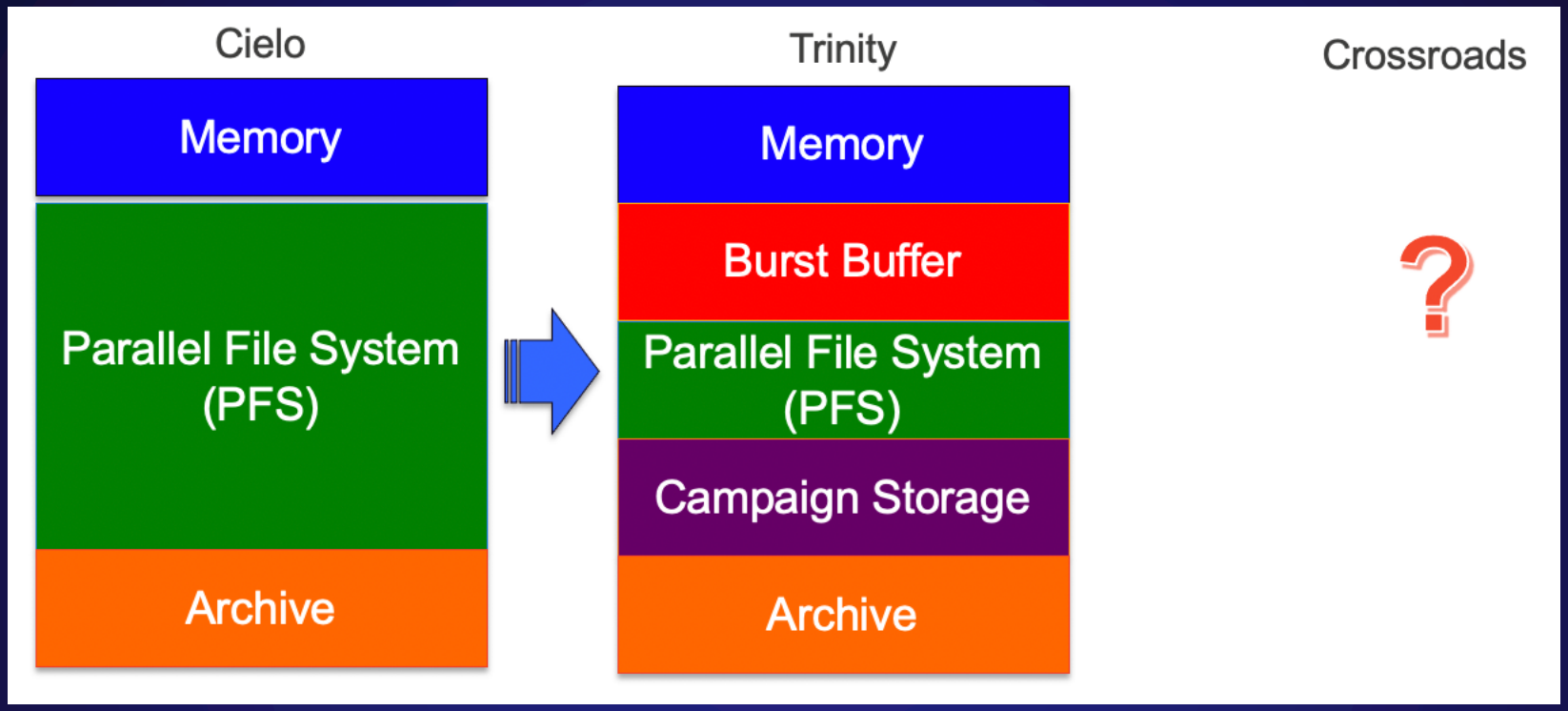
GOT MACHINES!

Fire: Penguin, 1390 teraflop/s
Ice: Penguin, 1390 teraflop/s
Cyclone: Penguin 1390 teraflop/s
Viewmaster3: HPE, Visualization
Thunder: Cray 167 node Arm TX-2
Moon: Appro 1600 systems testbed
Grizzly: Penguin, 1890 teraflop/s
Chicoma: HPE Olympus (Shasta) 3000 teraflops **NEW**
Badger: Penguin, 450 teraflop/s
Kodiak: Penguin GPU, 1806 teraflop/s
Snow: Penguin, 464 teraflop/s
Trinitite: Cray Haswell/KNL 384 teraflop/s
Capulin: Cray 175 node Arm TX-2
Moonlight: Appro, 488 teraflop/s open sys testbed
Lightshow: Appro, Visualization
Darwin: test bed for apps on architectures
Ising: Dwave going to 5000 qubits

Simple View of our Computing Environment



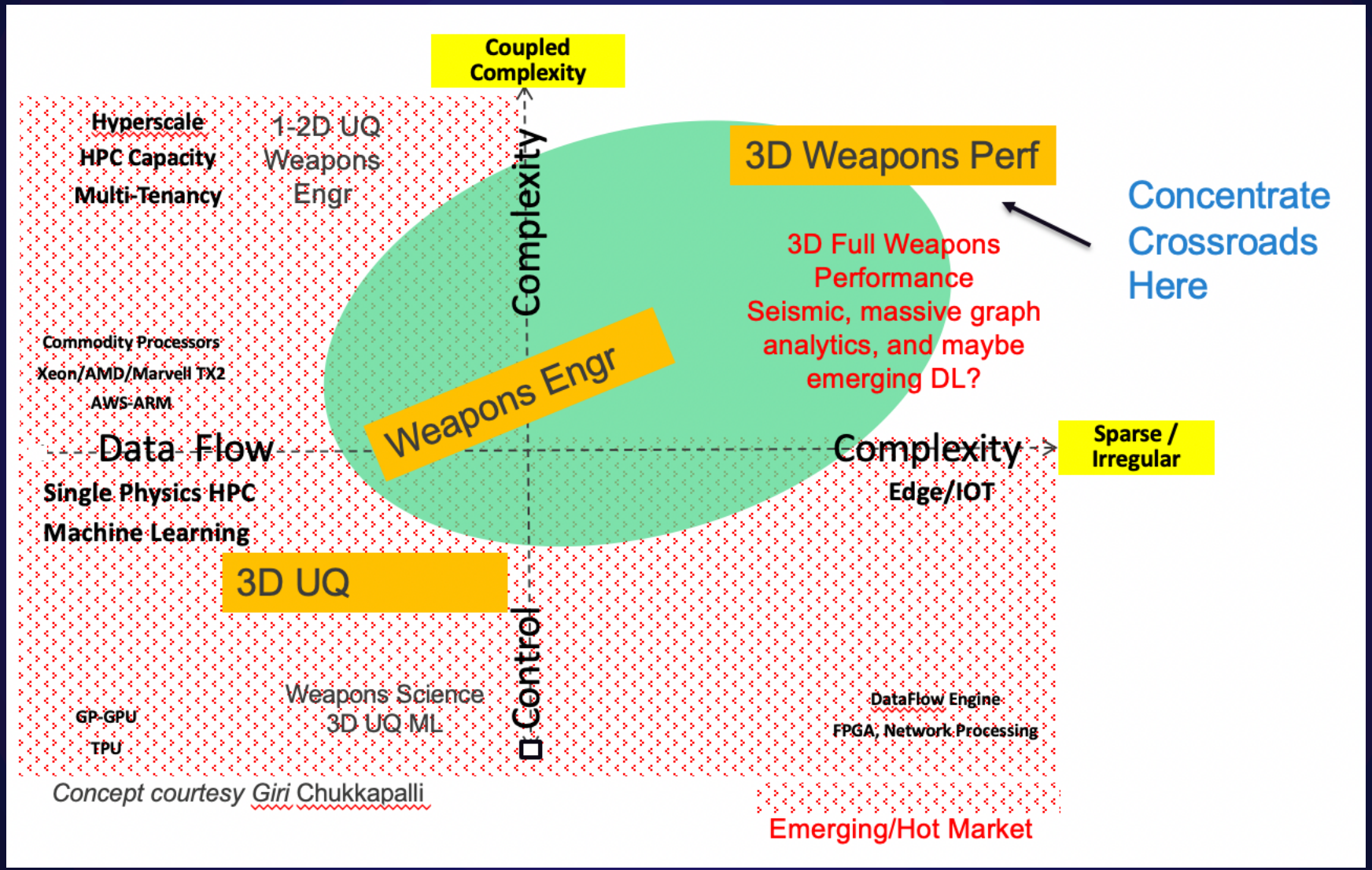
Tiers of Storage



- We don't want tiers but economically we must
- We HATE policy driven tiering – treat everyone equally badly doesn't work for a site that clearly has more important much larger/longer running work/users
- Our Campaign storage is working well and is extremely reliable with tiered erasure and much encoding and performs well

Crossroads
NNSA/ASC ATS-3

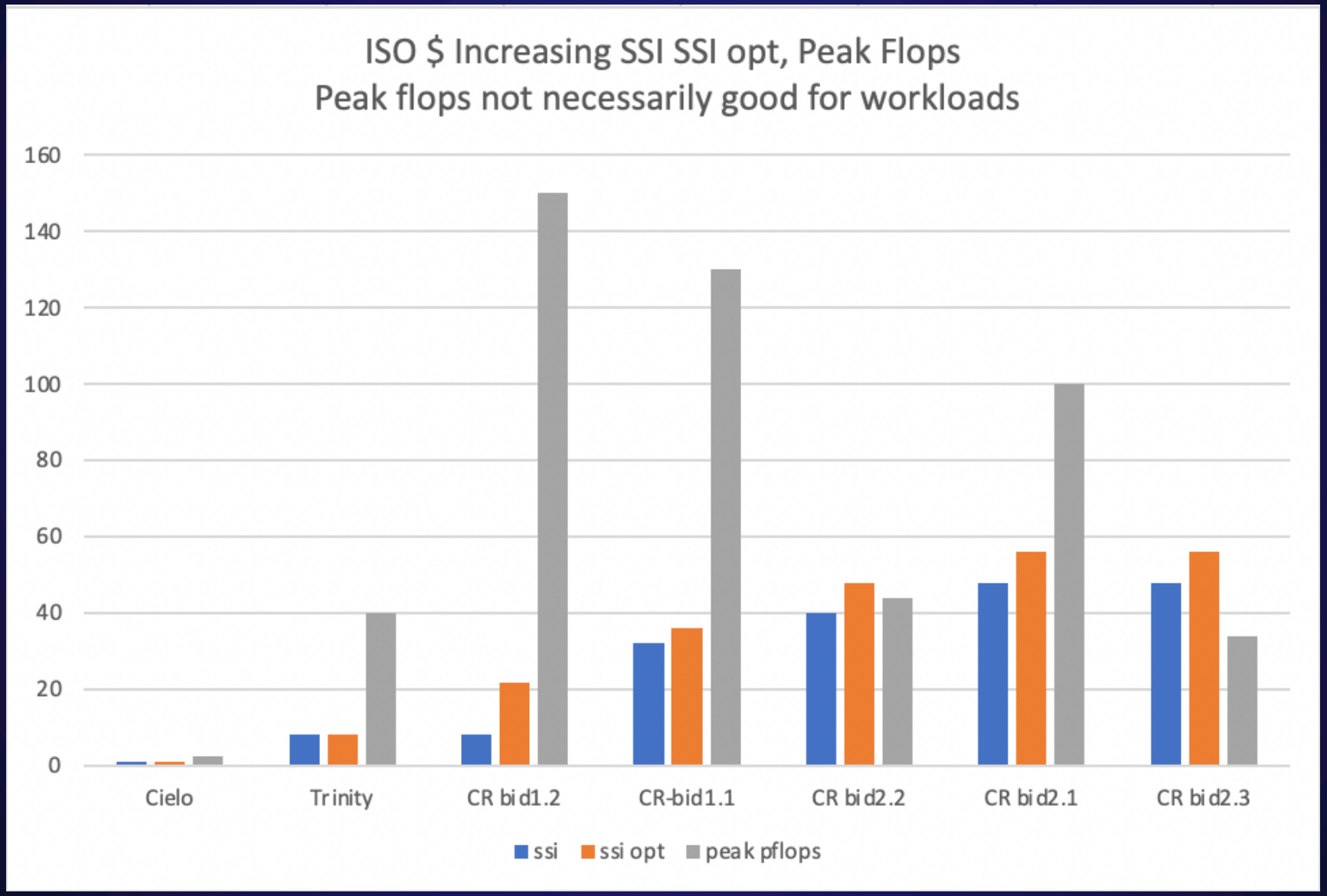
Crossroads Focus



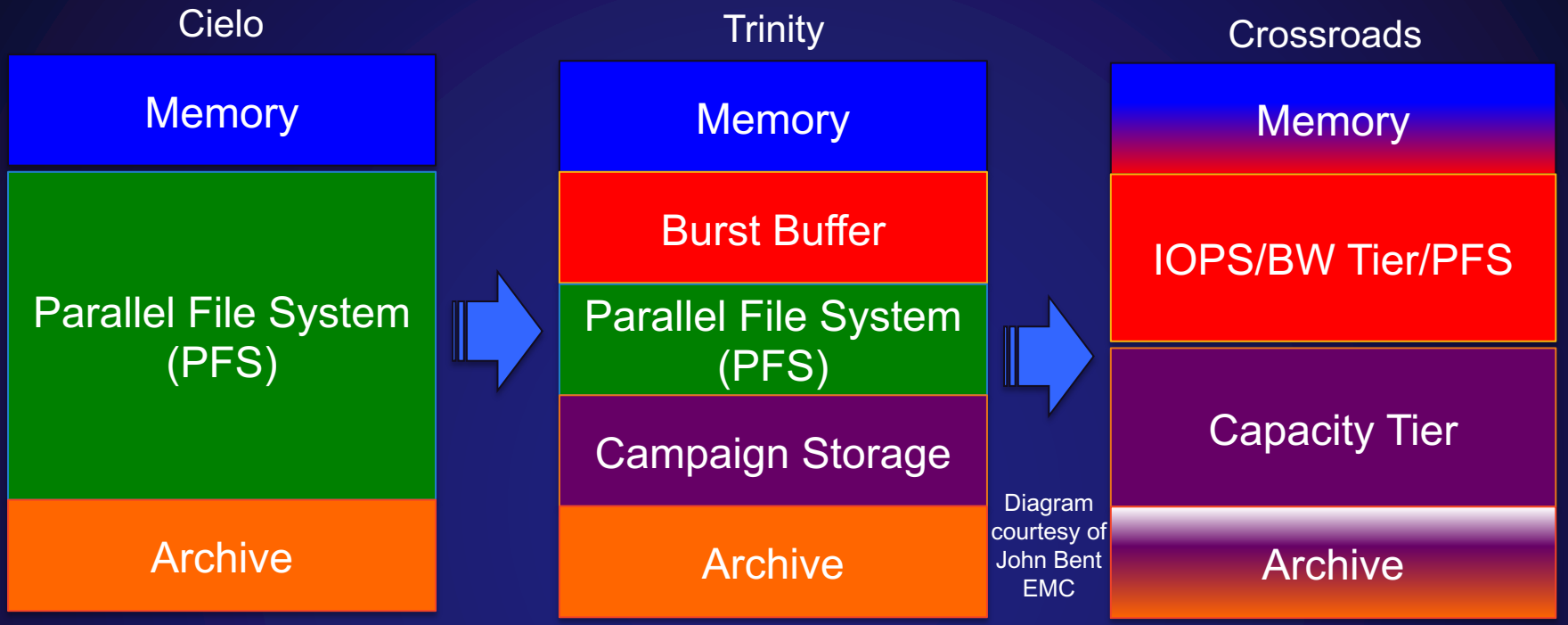
Crossroads V2 – RFP methodology (first step toward efficient computing direction for 3D Weapons Perf)

- **Minimum memory, max power/cooling/space**
- **Best value for Weapons Performance type Calculations**
- **NO FLOPS requirement specified in RFP!**
- **SSI (mini apps representing workload and workflow information about apps) (baseline and vendor modified mini apps)**
- **Most work through the system in its lifetime at LANL for lowest price wins**
- **Concentrate on:**
- **Performance efficiency** – The achieved performance of the application once enabled on the proposed **platform.**
  - Measure: improvement over previous platforms.
- **Workflow efficiency – The efficiency that a complete NNSA workflow executes on the proposed platform.**
  - Measure: predicted number of workflows that can be done in a platform's life
- **Porting efficiency – Large gains for each code/alg change**
  - Measure: scale/types of changes vendor makes to supplied mini-apps and the implications on overall code.

Flops or SSI (for Weapons Performance and other complex 3D/multi resolution/link scale/physics apps)



Crossroads Tiers of Storage



- **Campaign working very well - moving to be our primary capacity tier**
- **Economic allow move to flash only PFS on Crossroads**
- **Workflow driven staging between tiers/other storage**
- **We are contemplating moving the Archive from HPSS to Marchive (MarFS erasured tape) *** future *****

Crossroads Timeline

Crossroads Acceptance

Complete Delivery

Secure Production
Oct 3

CY 2021 | Q1 | Q2 | Q3 | Q4 | Q1 2022 | Q2 | Q3 | Q4 | CY 2022

Today

Center Of Excellence
- CoE

Crossroads Infrastructure Project (CIP)
- Construction — 229 days
- Project Complete — Feb 18

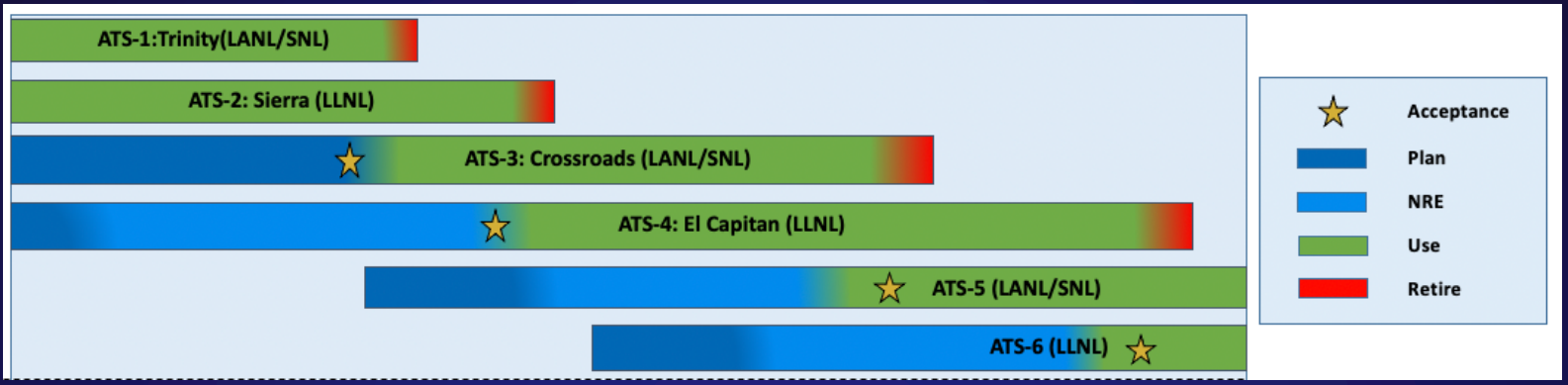
Build / Install
- Volume Parts — Mar 15
- 05/13 - 06/30 — Delivery — 34 days
- Acceptance — 44 days
- Transition to Secure — 20 days

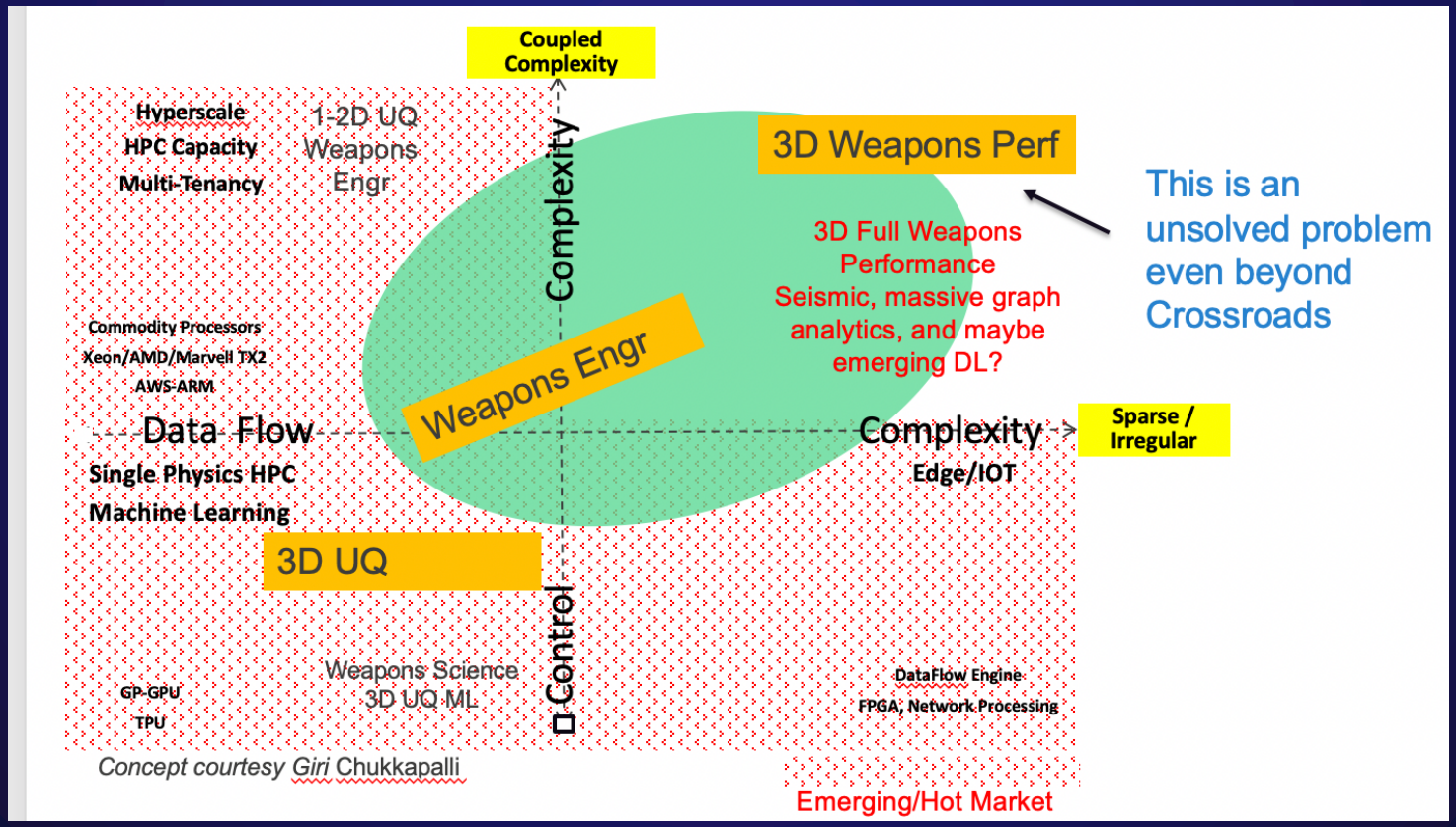
ASC Milestones
- System Readiness L2 — Sep 9
- Production Readiness L2 — Dec 23

Future Planning

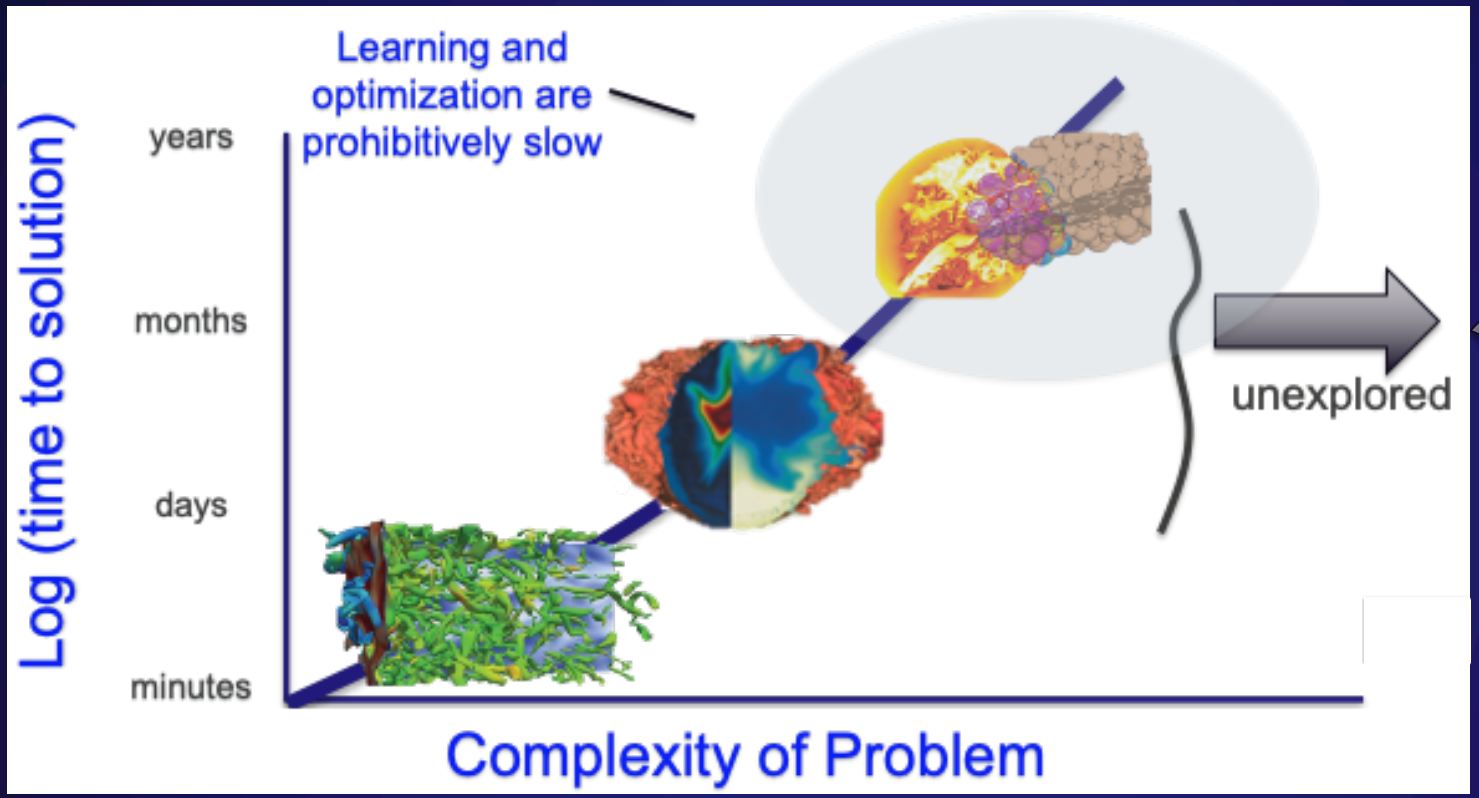
30X on our hardest problems to enable human learning!

Focus for the future





Our hardest and most important problem is to transform the study of the most complex systems

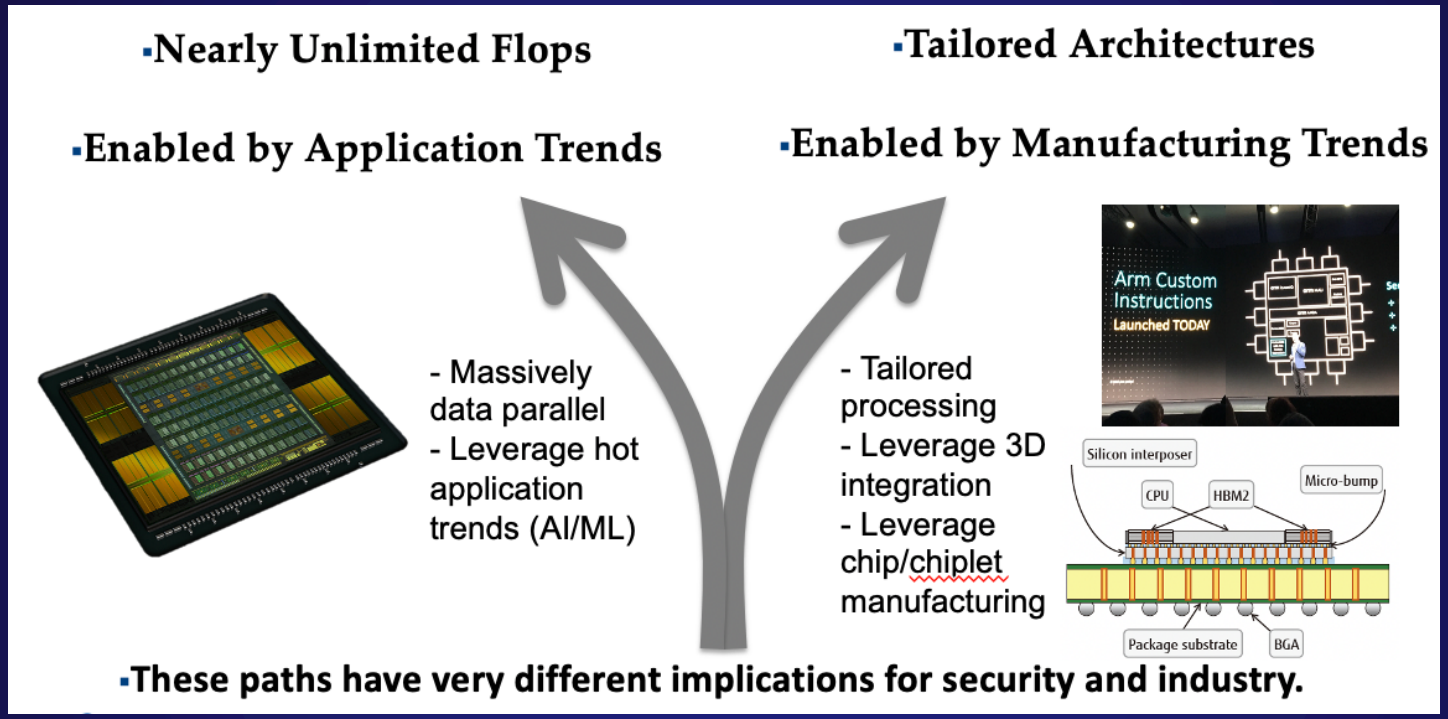


Concept courtesy Jason Pruet

These problems map least well to past and current computing architectures: coupled complexity and sparse / irregular (the dreaded upper right hand corner)

In order to shrink solution time from months to days (30X) we must address the long poles, coupled complexity and sparse / irregular

Architecture guided apps Or
Apps Guided Architectures – our Crossroads RFP
Experience would indicate the latter is right for LANL



- **Affordable 2.5D/3D integration/tailoring possible game changer**
  - Licensable IP (like Arm) is now somewhat affordable
- **Massive potentially leverageable industry directions**
  - Smart Nic direct from cpu/gpu
  - Smart Storage (NVME computational storage) direct from cpu/gpu
- **Focus next machines/collaborations/efforts on this direction!**

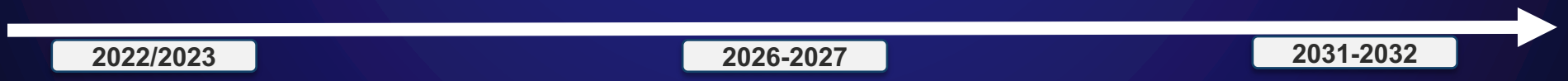
Systems path to >30X - pull technology into use

Crossroads
Production ATS workloads
(Begin the saga to more efficient computing on Weapons Performance)

Next Generation Stepping Stone
- LANL Inst use AI, ML, SIM
- LANL ASC use SIM, AI, ML
- Testbed future tech codesign / collaborations to guide 2026-2027
- Might contain mid life kicker for glimpsing into the future

ATS 5
- Larger machine than Crossroads, production ATS
- Code/workflow changes where necessary
- Major step toward 30X on most demanding simulation workloads employing fruits of codesign / collaboration / tailoring for sparse irregular /coupled complexity
- Demo sparse/complex AI/ DL

ATS 7
- Production ATS workloads
- Exceed 30X on light house apps
- Deploy next gen DL (sparse / complex)



2022/2023

2026-2027

2031-2032

Co-Design^2

Today's Co-design

Co-design^1 team membership:
- Science
- App/Alg
- CS Abstractions

Really more portability abstraction: Flexsi, Kokkos, Raja, to ease mapping to various architectures

Tomorrows Co-design

Co-design^2 team membership:
- Science
- App/Alg
- CS Abstractions
- Software to Hardware
- Hardware

Hardware is not a given, it has to change to meat the algorithms needs

Full Co-design science to hardware with crosscuts to look for common hardware customization opportunity

Co-design^2 team membership:
- Science
- App/Alg
- CS Abstractions
- Software to Hardware
- Hardware

Co-design^2 team membership:
- Science
- App/Alg
- CS Abstractions
- Software to Hardware
- Hardware

look for common opportunity

CS common

SW-HW common

HW common

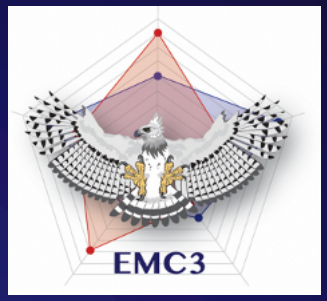
Like minded HPC technology producers/consumers banding together

Architectures Guided Apps → Apps Guided Architectures

Efficient Mission Centric Computing Consortium

EMC3

EMC^3 Areas of Interest



- **Networking**
  - Next Gen Network Requirements/Design
  - Apps/computing/programmability in network
- I/O & Storage futures
  - Data protection at extremes
  - Computation near storage
  - Data motion innovation
- Resilient Computing
  - Characterization/prediction
- **High Performance Data Analytics Systems**
  - Performance benchmarks/measurement of HPDA systems
- Inexact computing
  - Characterization for exploration
- **HPC Environments**
  - Launch, Run Time, Monitoring, Tools innovation

- Mapping Algorithms/Methods to architectures
- Processor/Memory Complex
  - Balanced Application focused
  - Power requirements
  - Scaling
- Benchmarking and Simulation of Systems
  - Use of Benchmarks/Simulation against codes to achieve balanced forward progress
- System SW
  - Next Gen Systems mgmt. (boot/launch/manage/etc.)
  - Leveraging container tech
- Application of ML/DL/AI to HPC

ATS3 ATS5 Ongoing

Thanks!

Join us in seeking backwards to efficient mission computing

Ultra-Scale Systems Research Center

Los Alamos NATIONAL LABORATORY EST. 1943

NNSA

HIGH PERFORMANCE COMPUTING

The Efficient Mission Centric Computing Consortium