Russian Supercomputing Days 2021

Active Memory Architecture: a non von Neumann Memory-Centric Paradigm for AI

Thomas Sterling

Director, AI Computing Systems Laboratory

Professor, Department of Intelligent Systems Engineering

School of Informatics, Computing, and Engineering

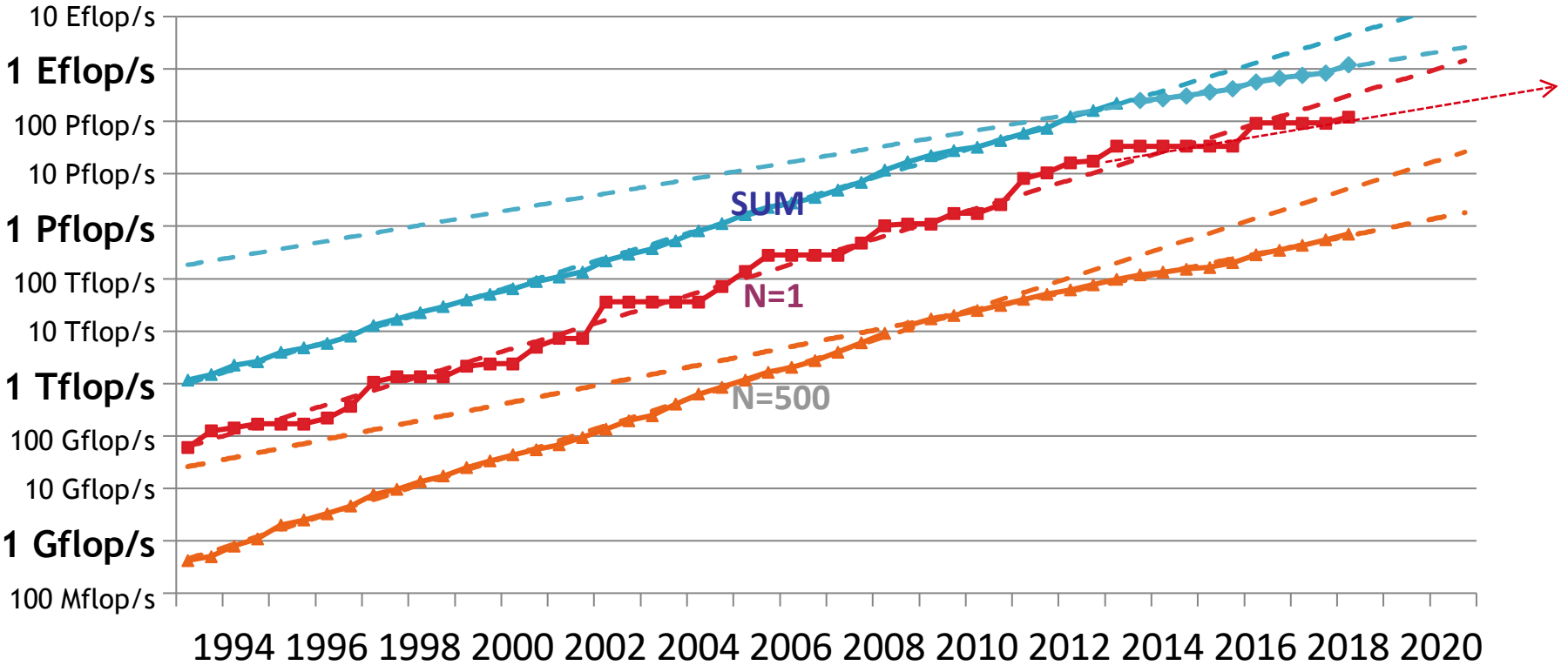
Indiana University Bloomington

September 28, 2021

At the Frontier of Exaflops

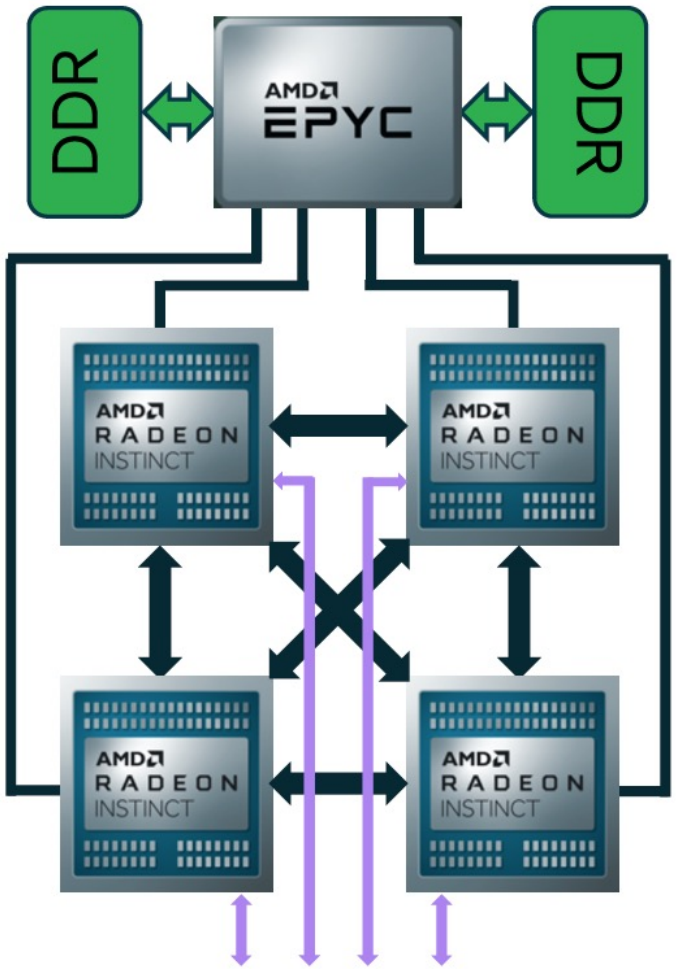
- 1st USA Exaflops R_{max} supercomputer – ~ 1.5 EFs
- Operated at DOE Oak Ridge National Laboratory (OLCF-5)
- Deployed late 2021 – early 2022
- Vendors: HPE Cray and AMD
- Power consumption: 30 MWatts
- Space: 104 racks
- Cost estimate: $600M

Projected Performance Development

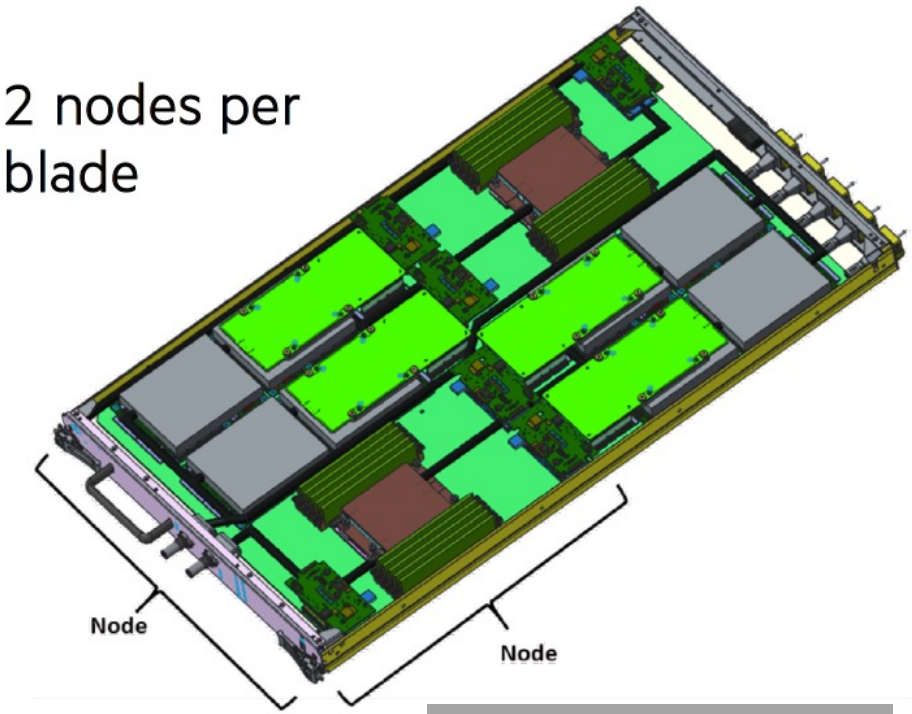


Courtesy of Erich Strohmaier

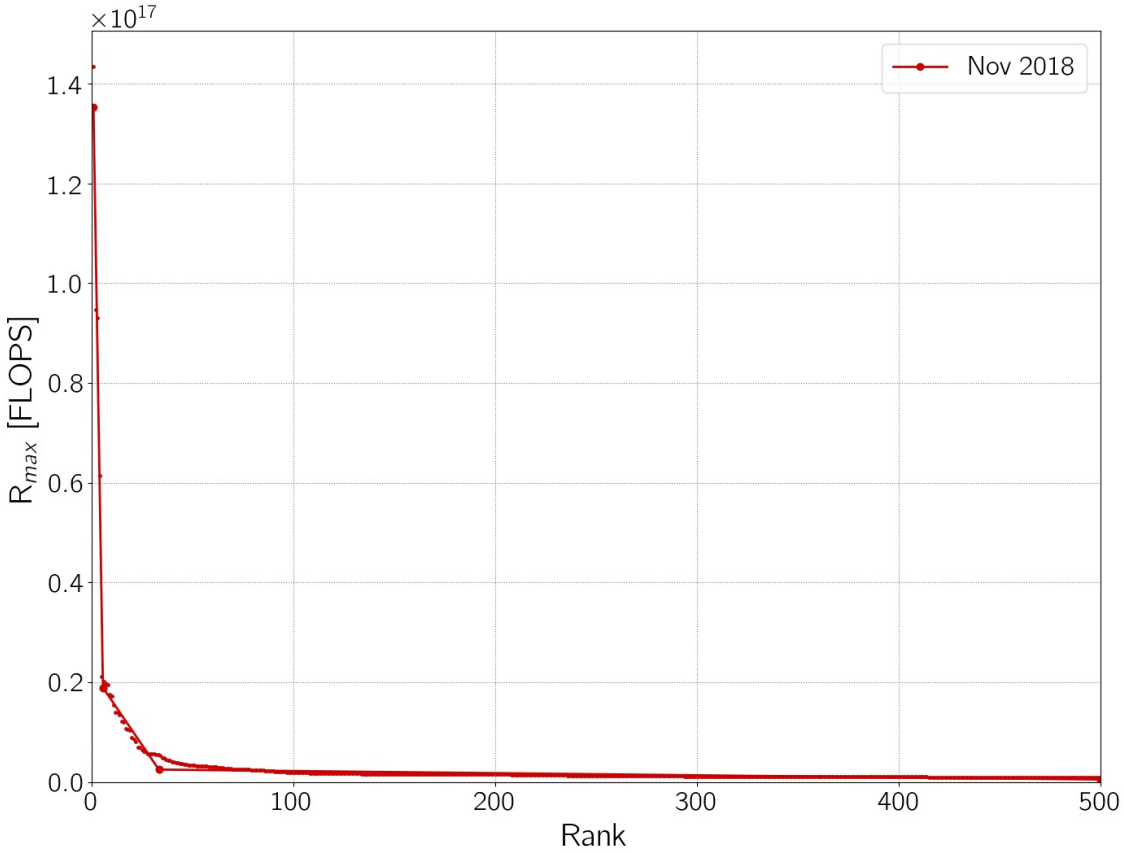
AMD GPU (ORNL)



2 nodes per blade

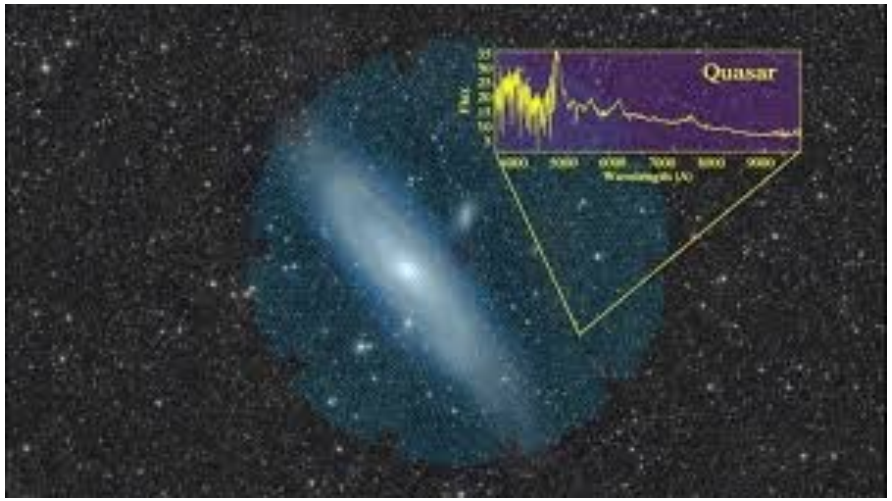


Three Worlds of Supercomputing



Heterogeneity for Mixed-Precision Analytics

- NERSC deploys Phase-1 of Perlmutter
  - Lawrence Berkeley National Laboratory (LBNL)
  - Approx. 60 Pflops R_{max} (DP FLOPS)
  - Approx. 4 Exaflops (16 bit floating point) for AI supercomputing
- 1,536 Nodes for Phase-1
  - 4 Nvidia A100 Tensor Core GPUs per node
  - 1 AMD EPYC AMD (“Milan”) processor per node
- Lustre filesystem
  - 35 Petabytes
  - 5 Terabytes/sec
- Applications
  - Controlled fusion
  - Climate modeling
  - AI for data analytics
  - Cosmology – dark energy
- Phase-2 will provide refresh with dual CPU only nodes



Technology Demands new Response

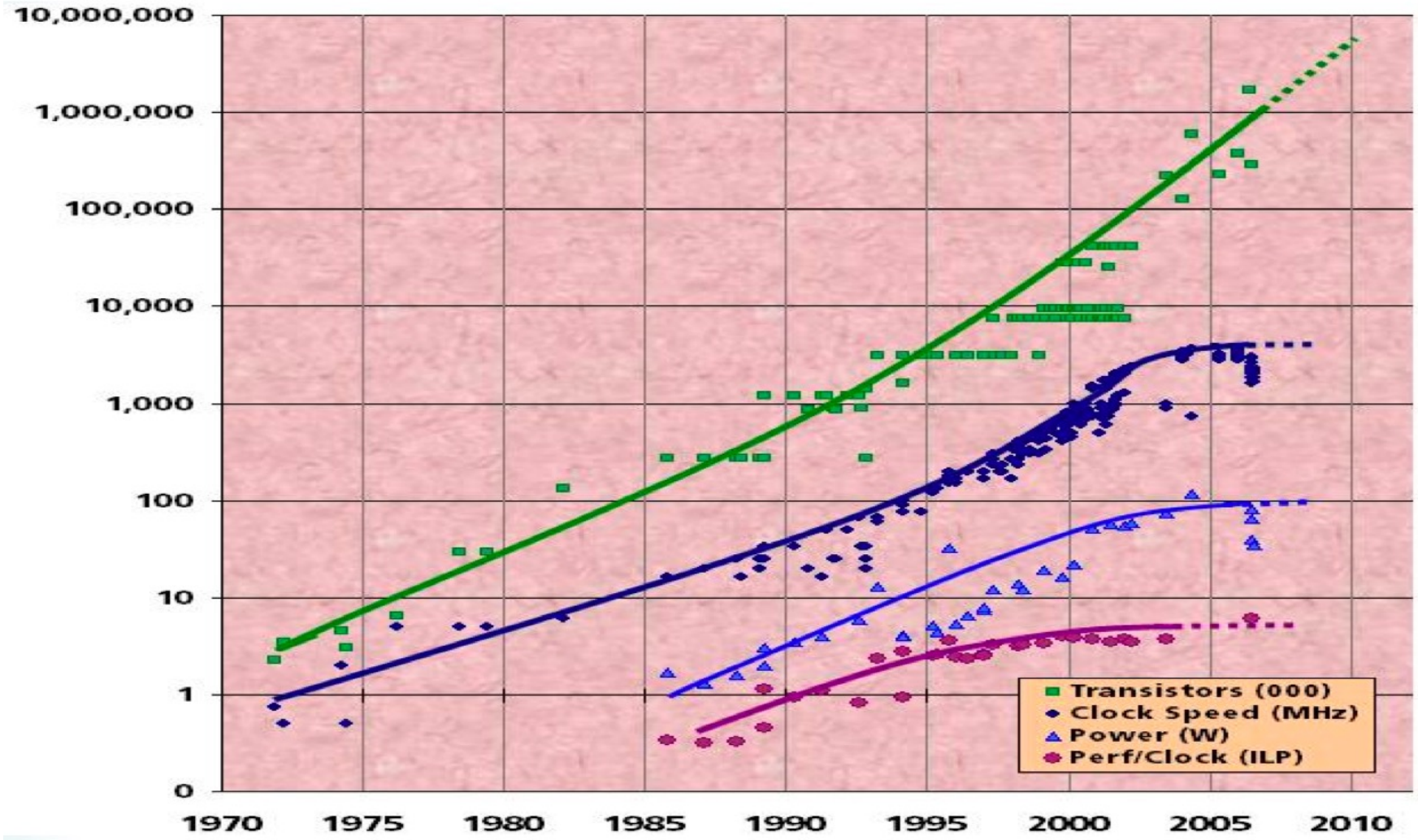


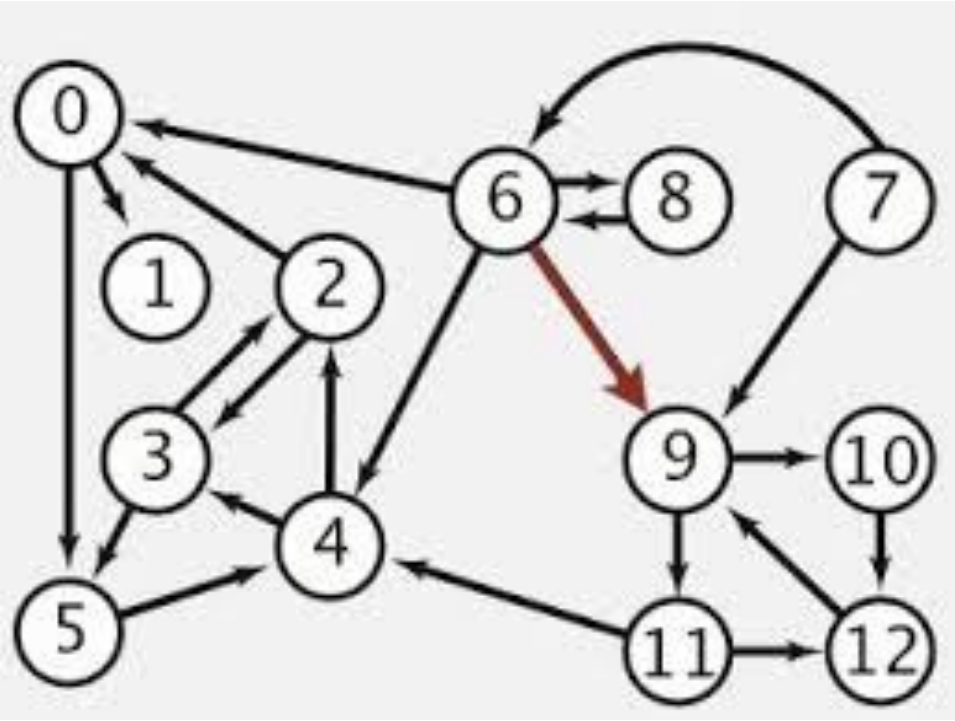
Figure courtesy of Kunle Olukotun, Lance Hammond, Herb Sutter, and Burton Smith

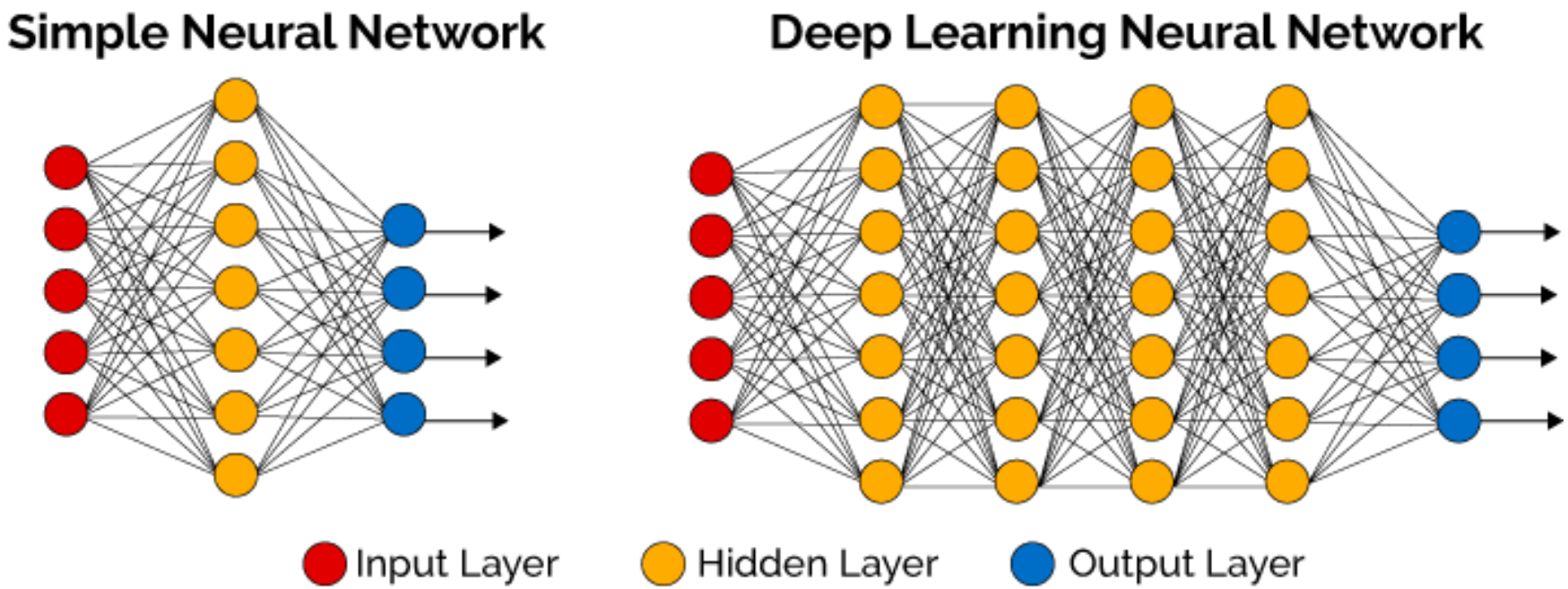
Von Neumann Architecture Barriers

- Optimization of ALU/FPU utilization
  - Presumes that FPU is most valuable resource
- Separation of compute logic and memory
  - Referred to as “von Neumann bottleneck”
  - Consequence of early disparate enabling technologies
- Sequential instruction issue
  - Constrains natural parallelism
- Sequential consistency memory model
  - Narrows ordering of global mutable side-effects
  - Imposes cache coherence mechanisms
- Registers of isolated local name space
  - Load/store operations
  - Logically insular
  - Causes register overflows
  - Fixed length

Graphs play Expanding Roles beyond just Data

- Represents Data Flow Execution
  - Static and Dynamic Dataflow
  - Systolic Arrays
  - Neuromorphic structures
  - Traces
- Search Structures
  - Scheduling
  - Planning
  - Optimization
  - Decision Making
  - NP hard computations
  - Heuristic-based Relaxation Tree
    - Memory is a constraint resource for partial state space representation
    - Vertices represent alternative choices with iterative Baysian confidence factors




Simple Neural Network
Deep Learning Neural Network
Input Layer
Hidden Layer
Output Layer

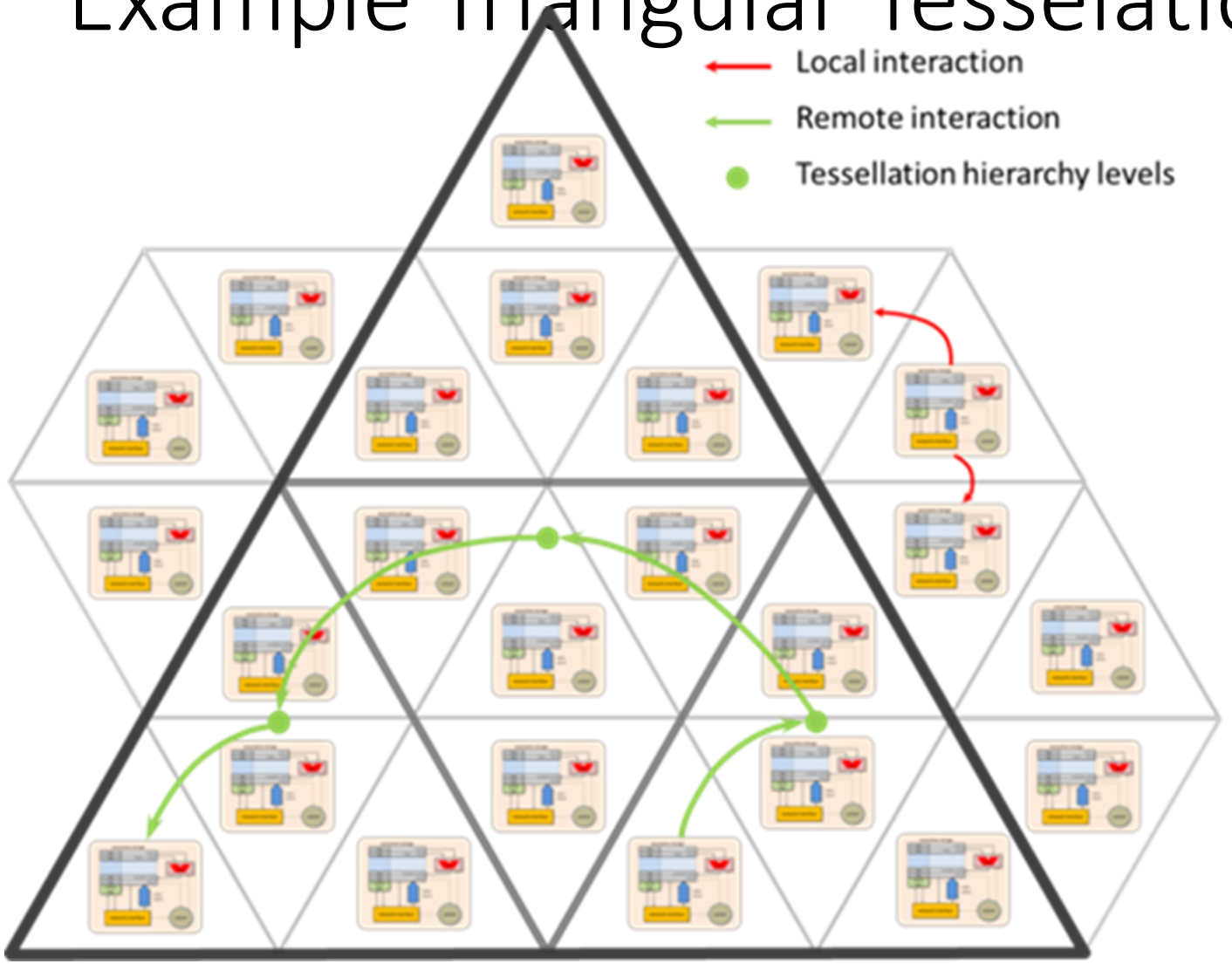
Challenges to Efficient Dynamic Graph Processing

- Merges data structure (vertices & edges) with direct operation flow control
  - Supported with innovation of hardware actions to minimize overhead
  - Extract parallelism through real-time system discovery
- Time-varying graph structures ill suited for sparse linear algebra
  - Sparse matrix mapping changes in many/all rows with even single edge variation
  - Preconditioning has to be reworked
- Undetermined rate at graph traversal
  - Distributed control to avoid fork-join constraints
  - Vertex local synchronization merged with data
  - Disambiguate order of arrival
- Transformation of graph topology
  - Distributed compound atomic operation
  - Addition or deletion of edge(s)
  - Addition or deletion of vertices

Non von Neumann Architecture Models

- Analog computing
  - Pre-von Neumann: Vannever Bush machines starting in 1930s
  - Technologies
- Cellular Automata
  - Invented by John von Neumann
  - Tessellated array of very simple cells
  - Cells comprise small state, rules, nearest neighbor access
- Data flow
  - Value-oriented (i.e., functional), message-driven (tokens)
  - Asynchronous, fine-grain
  - Static or Dynamic
- Neuro Morphic
  - Brain inspired, needs magic
  - Models neurons and synaptic junctions

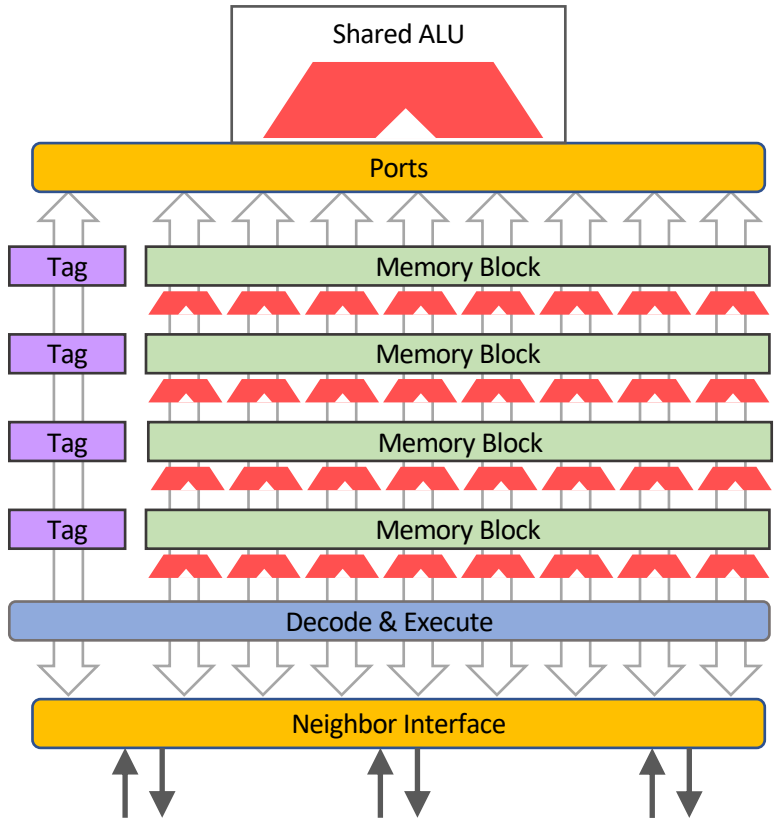
Example Triangular Tesselation



- Neighborhood routing
  - Wormhole routing
  - Dynamically modifiable on faults
  - About 1bit/hop
- Intra-die network
  - Hierarchy determined by tessellation boundaries
  - Fat tree – like
  - May be oversubscribed due to high-bandwidth adjacency traffic

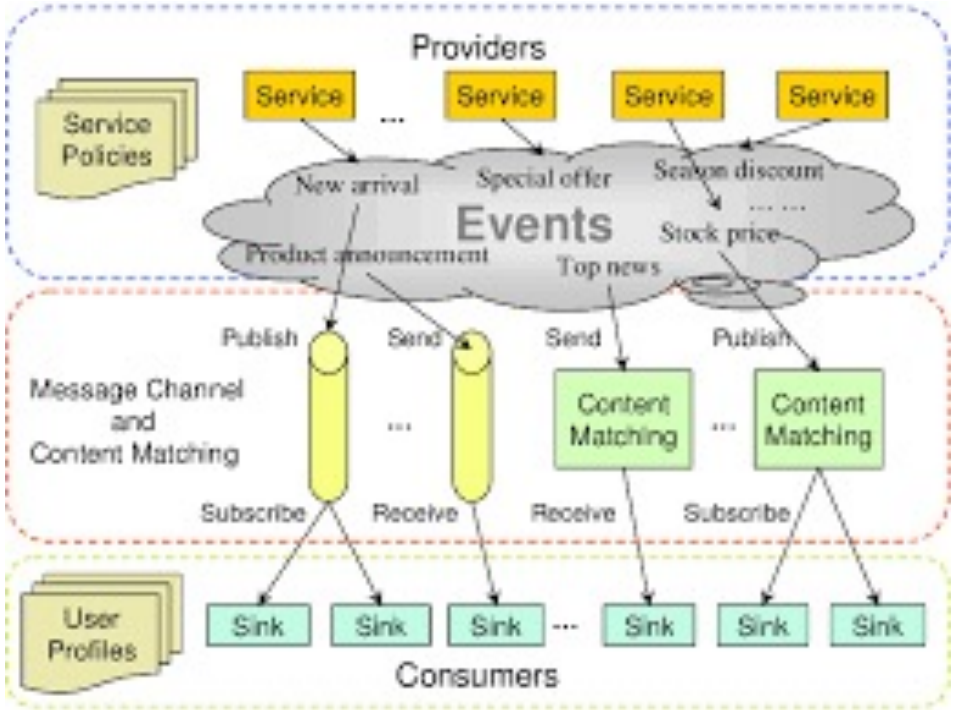
Active Memory Architecture – Compute Cell

- Compute-cell
  - Storage
  - Logic
  - Neighbor interface
- On-chip low-latency network
  - Hierarchical
- Socket
  - Stacked and unstacked versions
  - Fonton chip(s)
  - Stack includes DRAM and network driver chips
- Module
  - 1U enclosure
  - Four boards
- Global interconnect
  - On-board and inter-board
  - Electrical medium for short distance
  - High radix and fiber optics for machine scale

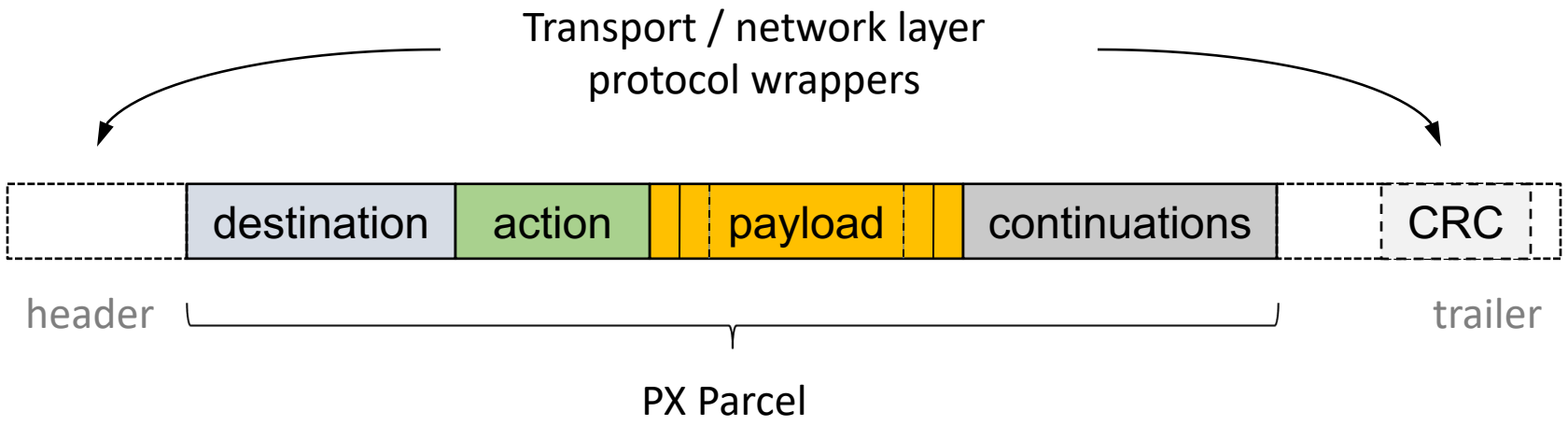


Motivation for Message-Driven Computation

- To achieve high scalability, efficiency, programmability
- To enable new models of computation
  - e.g., ParalleX
- To facilitate conventional models of computation
  - e.g., MPI
- Hide latency
  - Support overlap of communication with computation
  - Move work to data, not always data to work
- Work-queue model of computing
  - Segregate physical resource from abstract task
  - Circumvent blocking of resource utilization
- Support asynchrony of operation
- Maintain symmetry of semantics between synchronous and asynchronous operation



Parcel Structure



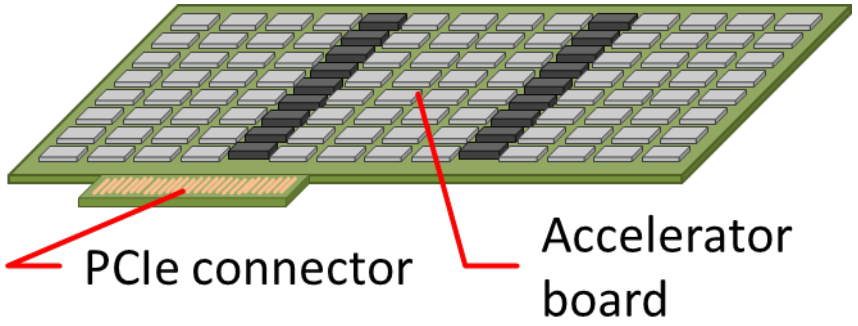
Parcels may utilize underlying communication protocol fields to minimize the message footprint (e.g. destination address, checksum)

Constraint-based Synchronization

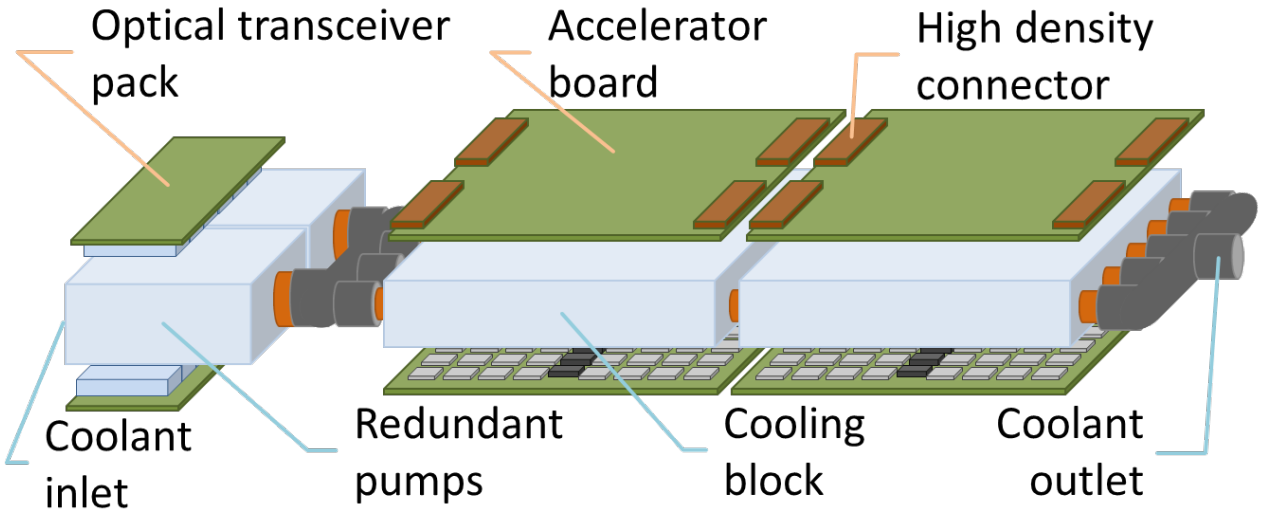
- Supports Dynamic-Adaptive Task Scheduling
- Declarative Semantics for Continuation of Execution
  - Defines conditions for work to be performed
  - Not imperative code by user
- Establishes Criteria for Task Instantiation
- Supports DAG flow control representation
- Examples:
  - Dataflow
  - Futures

Memory Accelerator

- GPU form factor
  - >100 Tops peak
  - PCIe connectivity
  - Unstacked chips
  - Discrete DRAM
  - Air-cooled



- Standalone module
  - 1 Petaflops peak
  - 1U rackmount
  - Stacked chips
  - Active liquid cooling

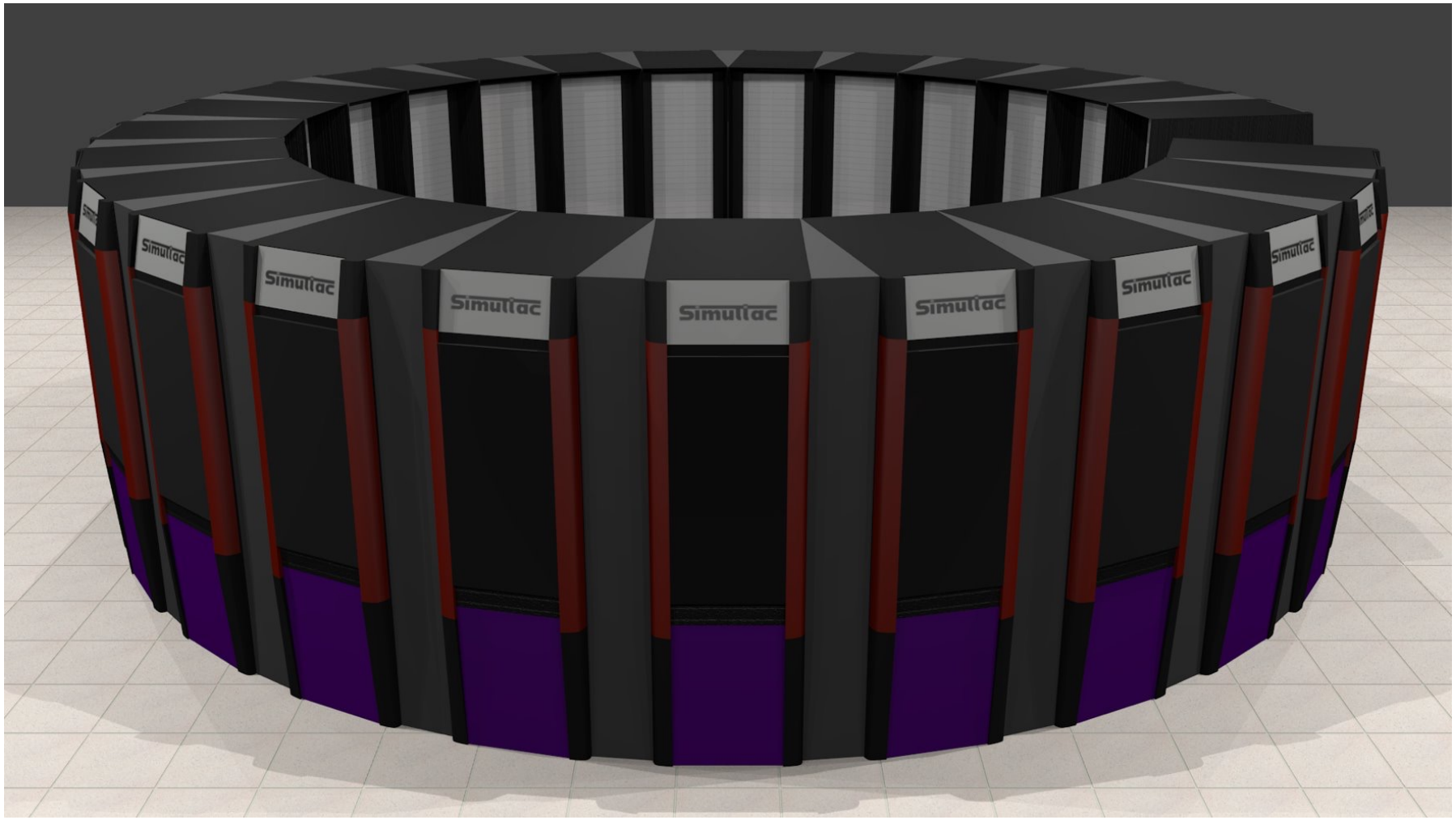


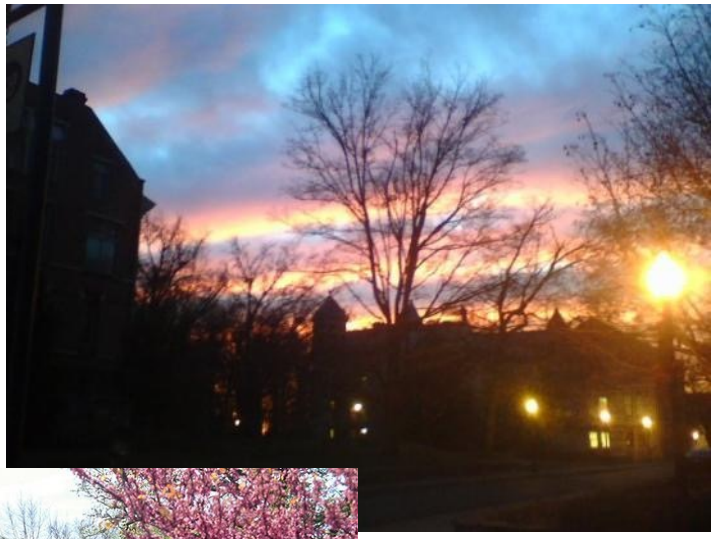
System Capacities and Capabilities

Level	C-Cells	Chips	Sockets	Boards	Modules
Chip	11K	1	-	-	-
Socket	43.8K	4	1	-	-
Board	16.3M	1,488	372	1	-
Module	65.2M	5,952	1,488	4	1
Rack	2.74G	250K	62.5K	168	42
System	65.7G	6M	1.5M	4,032	1,008

Level	Peak OPS	Peak Flops	Active memory [bytes]	DRAM [bytes]	Peak memory bandwidth [bytes/s]	Peak NN Bandwidth [bytes/s]
C-Cell	128M	16M	1K	-	3.07G	4.1G
Chip	1.4T	175G	11.2M	-	33.6T	44.9T
Socket	5.61T	701G	44.9M	1.07G	135T	179T
Board	2.09P	261T	16.7G	399G	50.1P	66.8P
Module	8.35P	1.04P	66.8G	1.6T	200P	267P
Rack	351P	43.8P	2.8T	67.1T	8.41E	11.2E
System	8.41E	1.05E	67.3T	1.61P	202E	269E

Exascale (DP FLOPS) with today's technology





See you (Hopefully) in Frankfurt in 2022