Scaling Up as well as Scaling Out

Jamon Bowen

Pete Jarvis

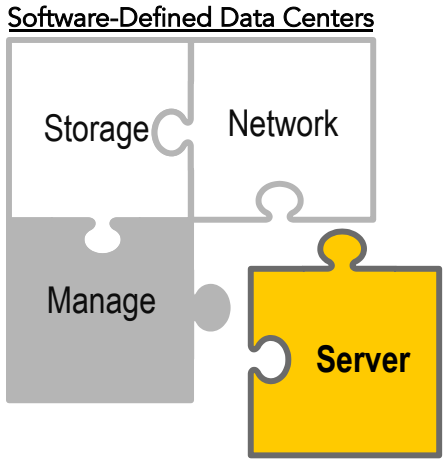
The Challenge

- The University of Texas at San Antonio (UTSA), a premier research university with strong academic and research programs in cyber security, cloud computing, computer analytics and big data.
  - Problem:
    - Researchers need access to on demand systems that run applications without requiring rewriting as parallel MPI jobs.
  - UTSA Objective:
    - Flexible and granular control to scale and re-provision resources on demand.

What If Servers were Software-Defined?

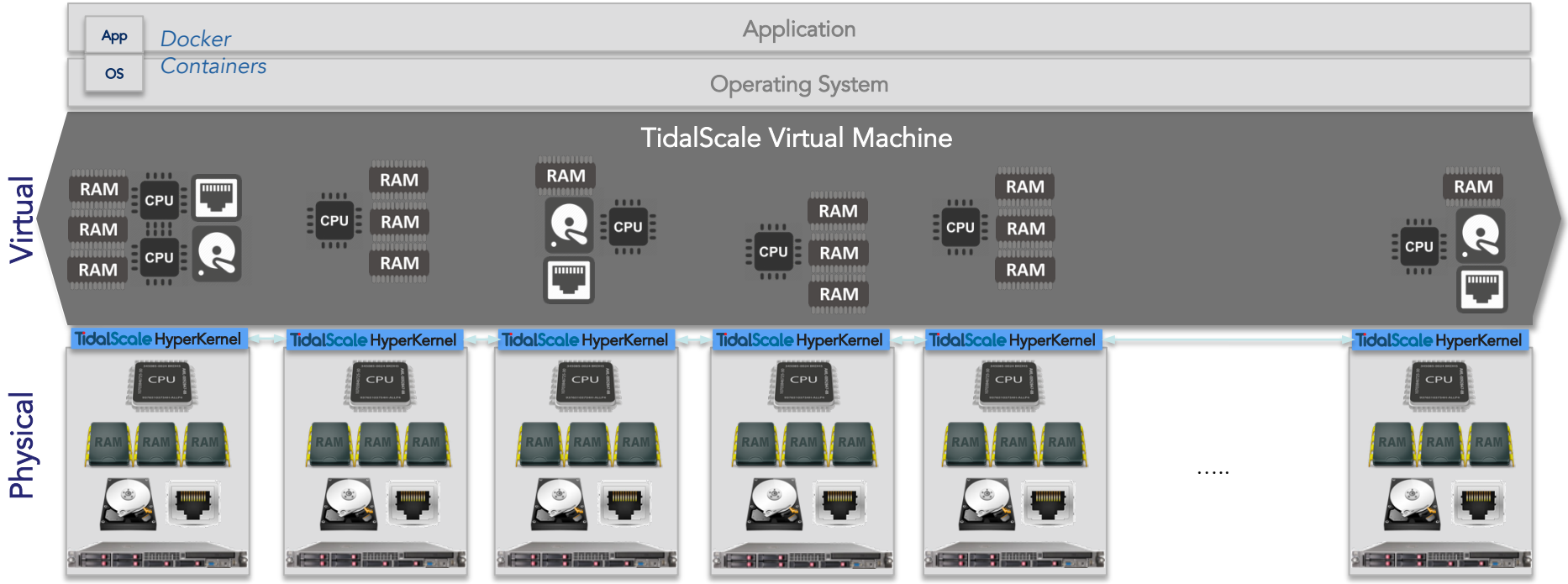
Flash Memory Summit

- In-memory performance at scale
- As many cores as needed
- Self optimizing
- Everything just works
- Uses standard hardware



Flash Memory Summit

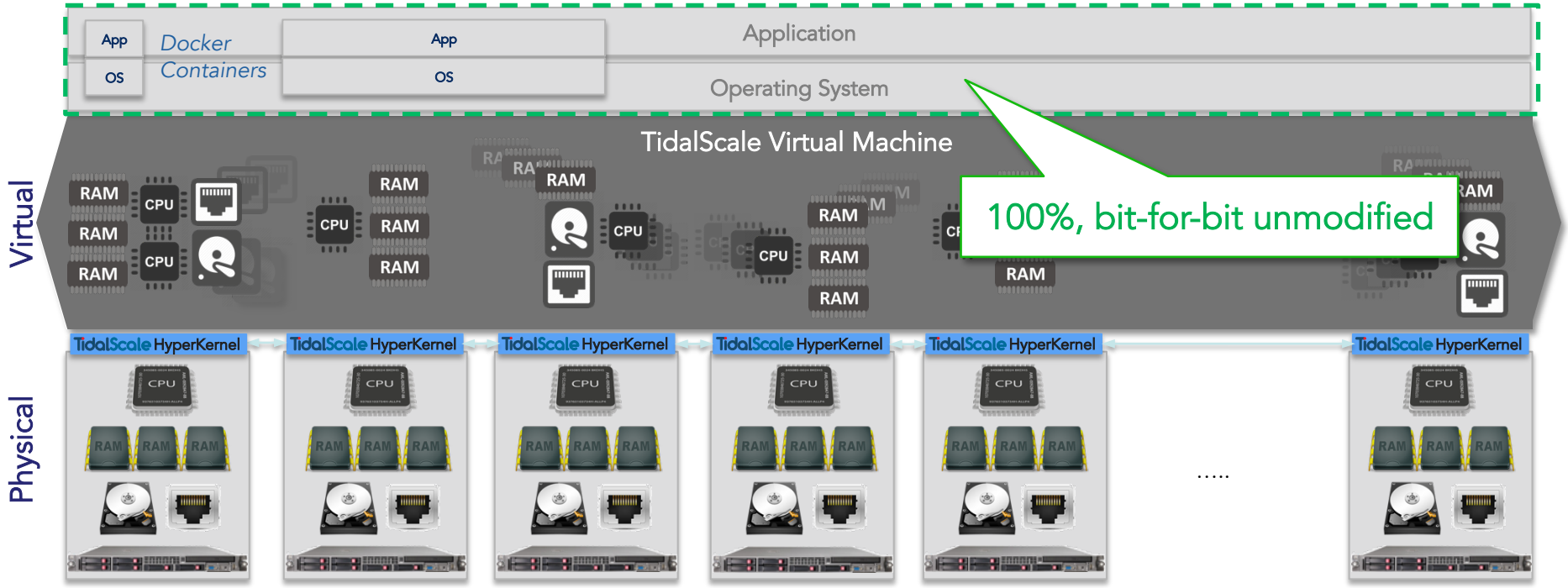
Solution: TidalScale Software-Defined Servers



Using patented Machine Learning algorithms, *mobilized* virtual resources flow across the physical resources to achieve optimal locality.

Flash Memory Summit

Solution: TidalScale Software-Defined Servers



Using patented Machine Learning algorithms, *mobilized* virtual resources flow across the physical resources to achieve optimal locality.

The Memory Cliff Hierarchy in Human Terms

Operation	Processing Latency	In Human Terms
1 CPU Cycle	0.3 ns	1 sec
L1 Cache	1 ns	3 sec
L2 Cache	3 ns	9 sec
L3 Cache	13 ns	45 sec
DRAM	50 ns	3 minutes
Memory over Ethernet	3 µs	3 hours
CPU Context Transfer	6 µs	6 hours

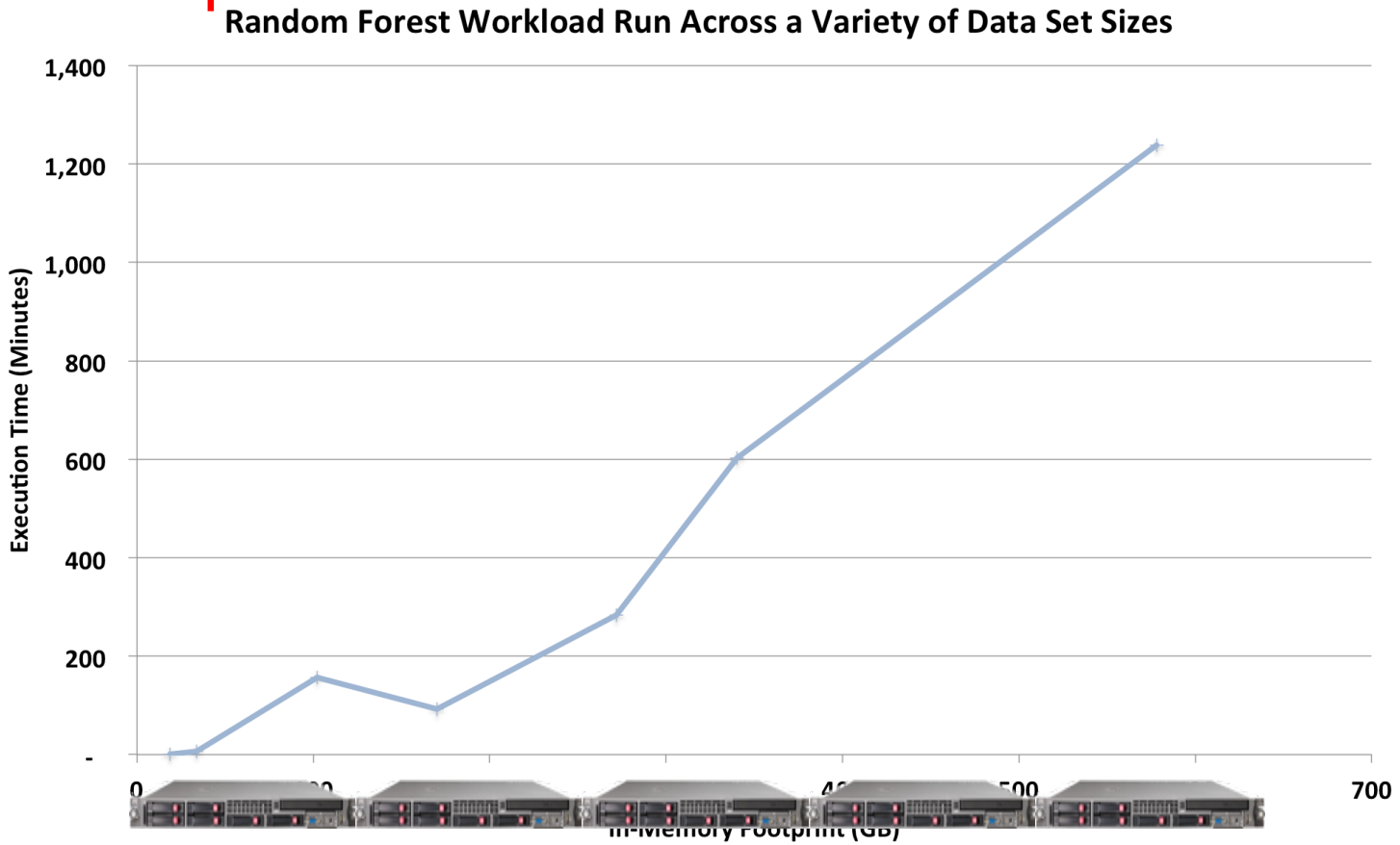
Memory Cliff

NVMe Flash	150 µs	6 days
Flash Array	1 ms	6 weeks
Internet: San Francisco to New York	40 ms	4 years
Internet: San Francisco to Australia	180 ms	19 years
TCP packet retransmit	2 s	211 years

TidalScale

Open Source R - Random Forest

Flash



Complete Benchmark Runtime

- No sharding
- No code changes
- In-memory performance across 5 nodes
- 240x to 550x faster