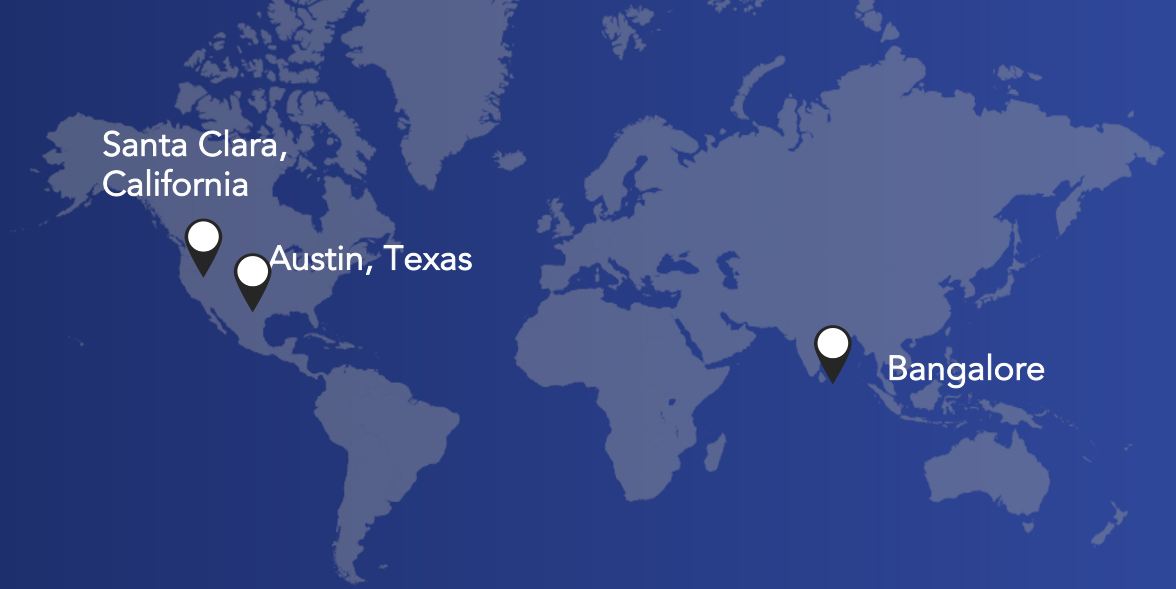
Elastics.
CLOUD

CXL: Saving Billions of Dollars with Shared Memory Pools and Lower Latency

Elastics.cloud Profile



Founded in **2020**

1st to demonstrate symmetric memory pooling

5 patents filed

Over **1500+** total years of product development experience

60+ engineers worldwide

Founders

George Apostol

CEO & Founder

- 35 years of experience designing system-on-chip (SoC), hardware, software, and systems.
- Leadership and executive roles at Xerox/PARC, Sun, SGI, LSI Logic, Exar, Samsung, TiVo, BRECIS.
- Holds several patents for interconnect and interface design .

Shreyas Shah

CTO, Chief Scientist & Founder

- 25 years of experience in the design of semiconductor, system design, and architecture in fields of computing, networking, storage technologies, virtualization and Flash based storage.
- Over 15 patents issued and numerous pending.

System Performance Challenges

System Demands

Exponentially increasing financial data

Increased processing due to AI/ML

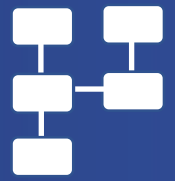
+

PCIe Higher Latency

PCIe stuck at latency of 1 µs per hop

+

Architectural Inefficiency



Excessive copying of data, inefficient utilization

=

Trading Inefficiencies

Every nanosecond of tick-to-trade latency equates to $Millions

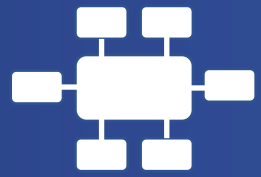
+

CXL Reduces Latency

Interconnect latency reduced by over 50%

+

Composable Architecture



Remove bottlenecks, add memory and cache capabilities

=

Optimize Trading, Lower TCO

Maximize profits, improve utilization

TCO: Memory and Power Cost Savings with Memory Pooling

	Without Pooling	With Pooling
Near CPU Socket Memory (6,800 GB/Server = 1/3 of the memory)	$217,600	$217,600
Pool-able Memory (6,800 GB/Server = 1/3 of the memory)	$217,600	$261,120* *20% CXL premium
Stranded Memory (6,800 GB/Server = 1/3 of the memory)	$217,600	$0
Power	$60,000	$40,000
Total Cost	$712,800	$518,720

$194,080* Savings
*Based on a 40 server rack

$80B memory and power cost savings
(based on 400K racks/year)
- 512 GB/server
- 40 server/rack
- 20 TB total memory per rack
- $32/GB

Better Utilization;
Increased Capacity;
Improved Performance;
Lower TCO

Elastics.cloud Confidential

Elastics.cloud

CXL Improves the Bottom Line in Two Ways

Increased Profitability

- Reduce latency for individual connections
- Reduce overall system latency
- Enable larger data sets per server
- Allow for more streamlined systems architecture, reducing bottlenecks in CPU access

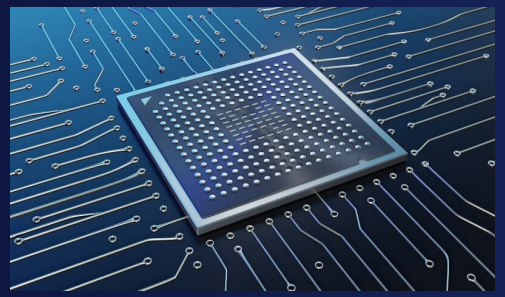
Reduced TCO

- Reduce stranded memory costs
- Higher memory/resource density in existing server footprint
- Better utilization of scarce real data center space close to exchange
- Reduce number of server skus based on memory footprint

Elastics.cloud Composability Solutions

Current

AMD/Intel FPGA
running Elastics.cloud IP

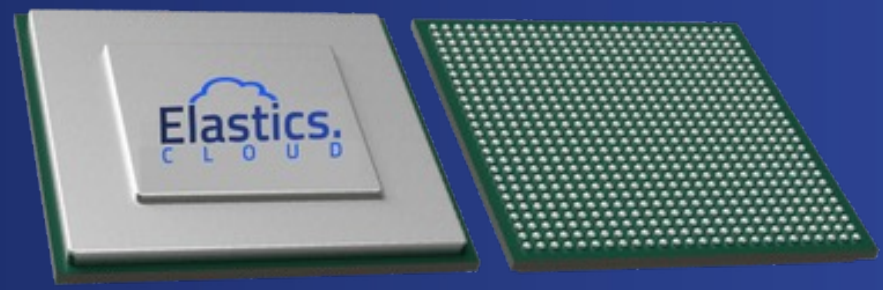


Elastics.cloud IP

CXL1.1/CXL 2.0
Memory Expansion
Memory Pooling
Accelerator Pooling

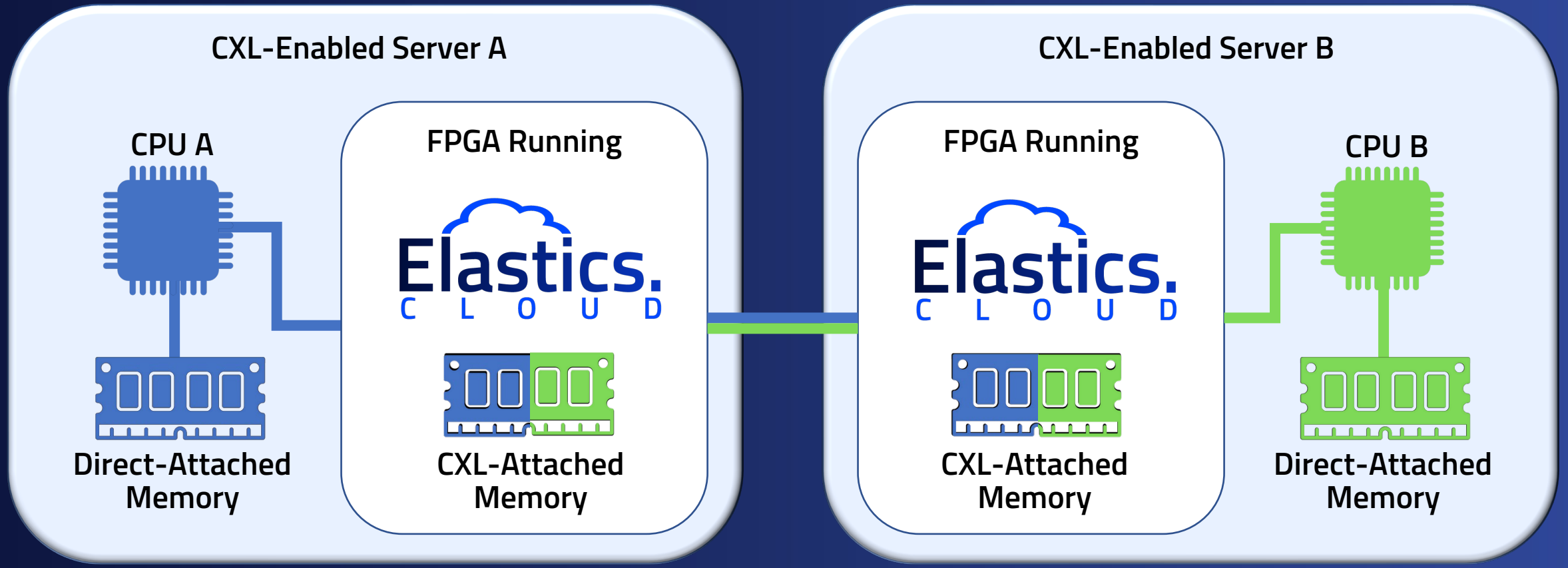
Future

Elastics.cloud SSoC
(Switch System on Chip)



CXL 2.x
Memory Expansion
Memory Pooling
Accelerator Pooling
Full Composability
Added Features & Intelligence

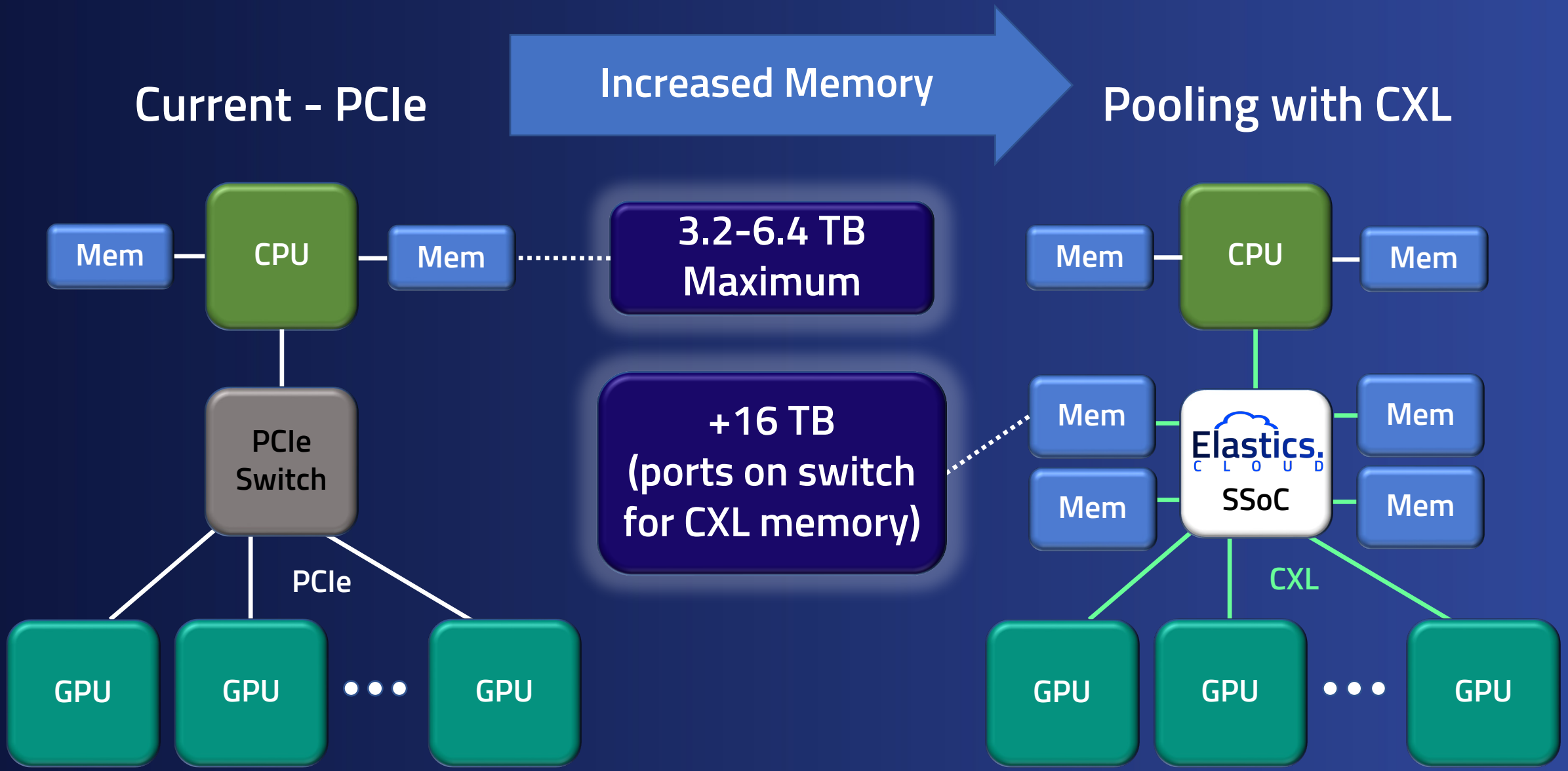
Elastics.cloud First to Demonstrate Symmetric Memory Pooling



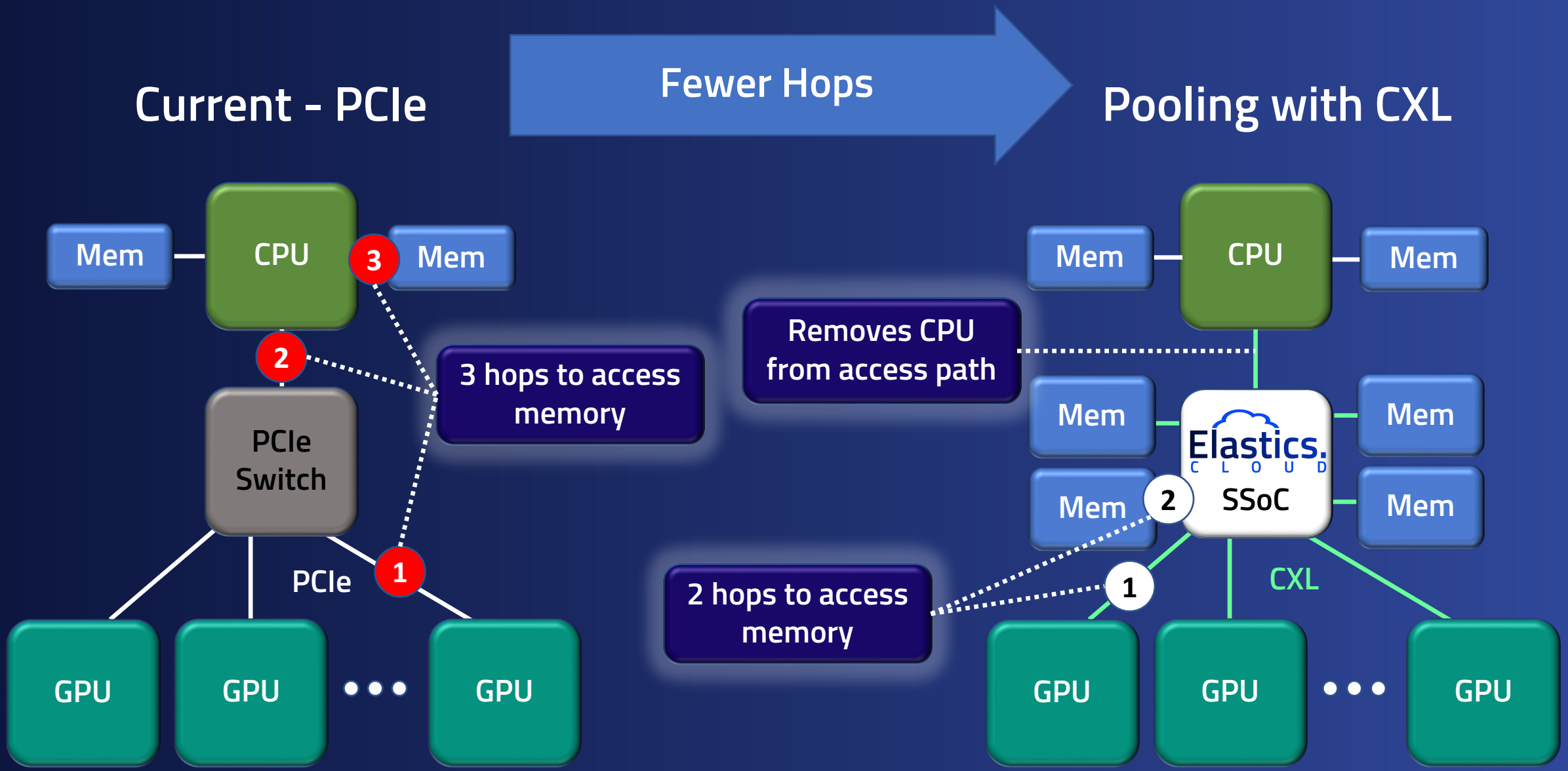
CPU A Accesses:
Server A Direct-Attached
Server A CXL-Attached
Server B CXL-Attached

CPU B Accesses:
Server B Direct-Attached
Server B CXL-Attached
Server A CXL-Attached

Pooling Architecture with Elastics.cloud SSoC

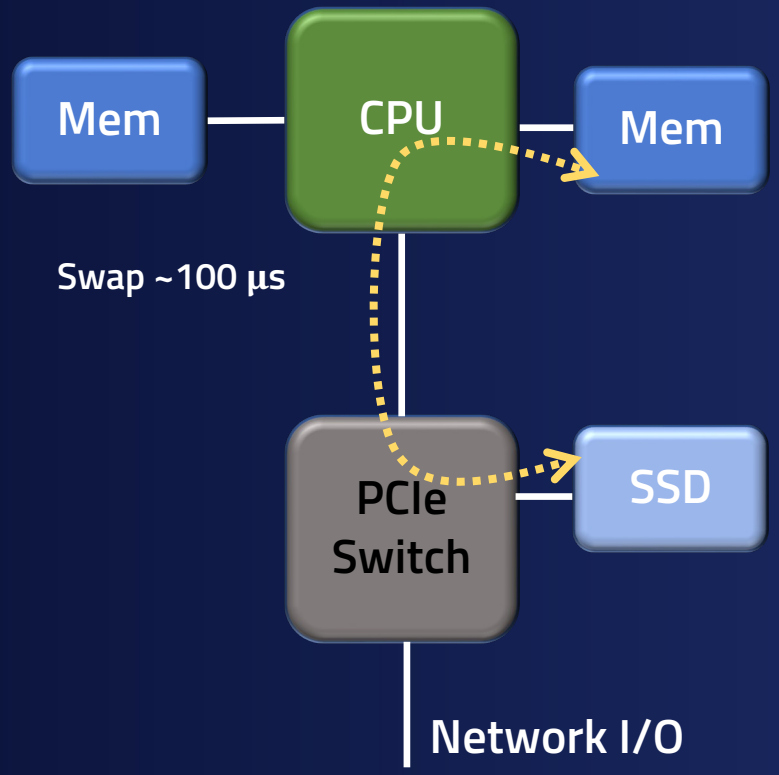


Pooling Architecture with Elastics.cloud SSoC



Memory Expansion & Pooling Performance Improvements

Current - PCIe

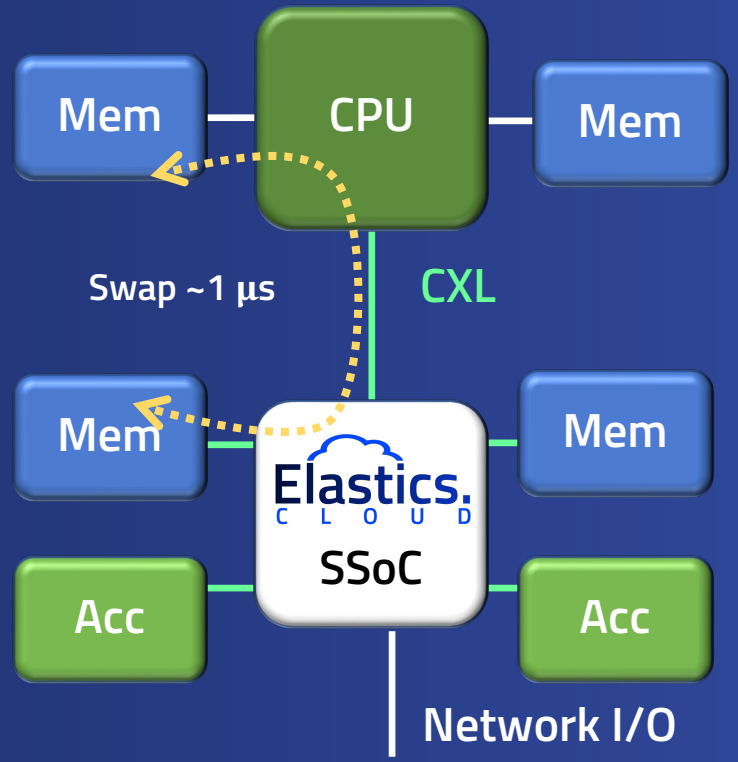


Lower Latency and Ops/Sec

Hardware swap latency: ~100 µs to ~1 µs

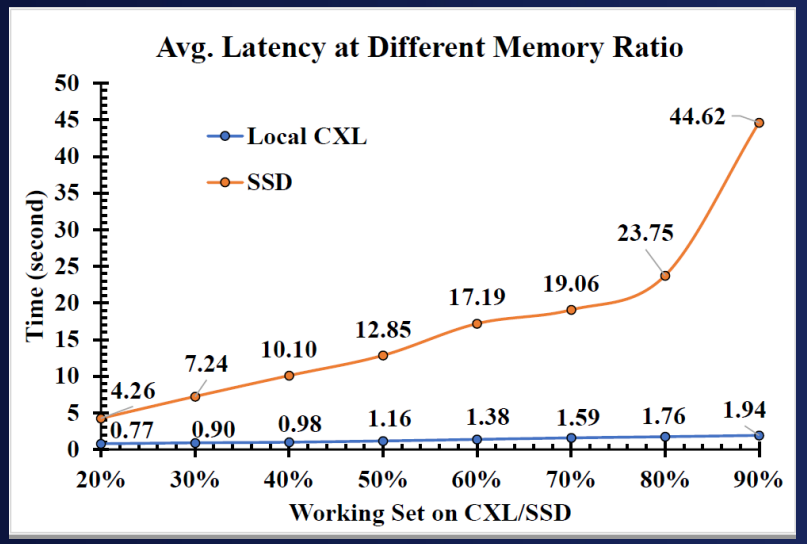
DB App: up to ~23x more ops/sec, 22x lower latency

Memory Expansion with CXL Peer to Peer Transfers



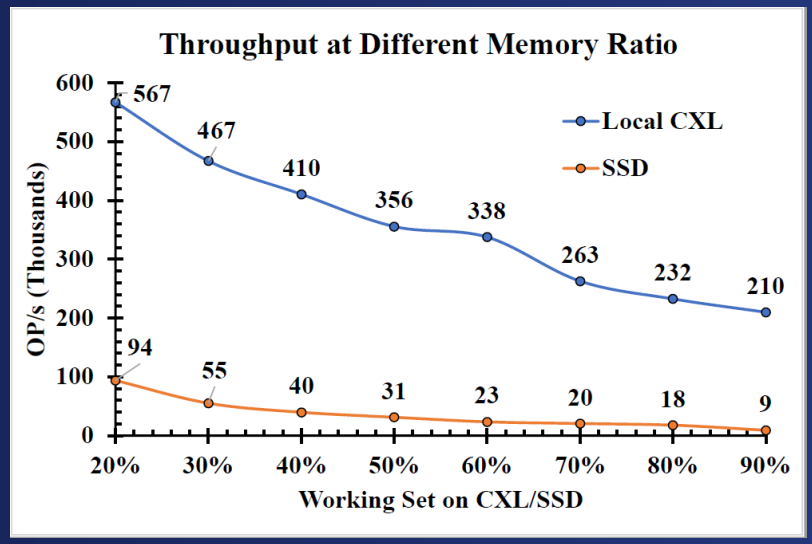
Redis Labs DB: Swap from SSD to CXL-Attached Memory

Latency
up to 22x lower



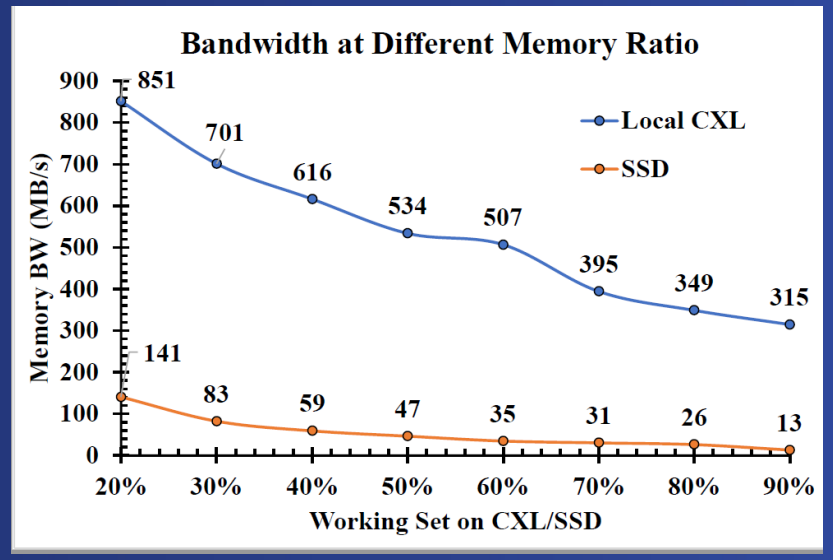
Graph demonstrates average latency (seconds)/operation

Ops/sec
up to 23x higher



Graphs demonstrates operations/sec

Bandwidth
up to 24x higher



Graphs demonstrates memory bandwidth MB/sec

Datasets increase in size, swap from SSD to CXL - 20% to 90%

Conclusions

1. CXL protocol will provide improved performance and lower latency than PCIe

2. System level utilization and lower TCO can be realized with new CXL-based architectures

3. Advantages are achieved above what the connectivity provides for systems applications like in-memory databases

4. Resource sharing (i.e. memory, accelerators, storage) improves utilization and performance

5. Elastics.cloud patent-pending solutions provide additional systems-level features further improving performance and lowering TCO at scale

Elastics.
CLOUD

Interested in a demo of our composable memory solution?

Contact sales@elastics.cloud