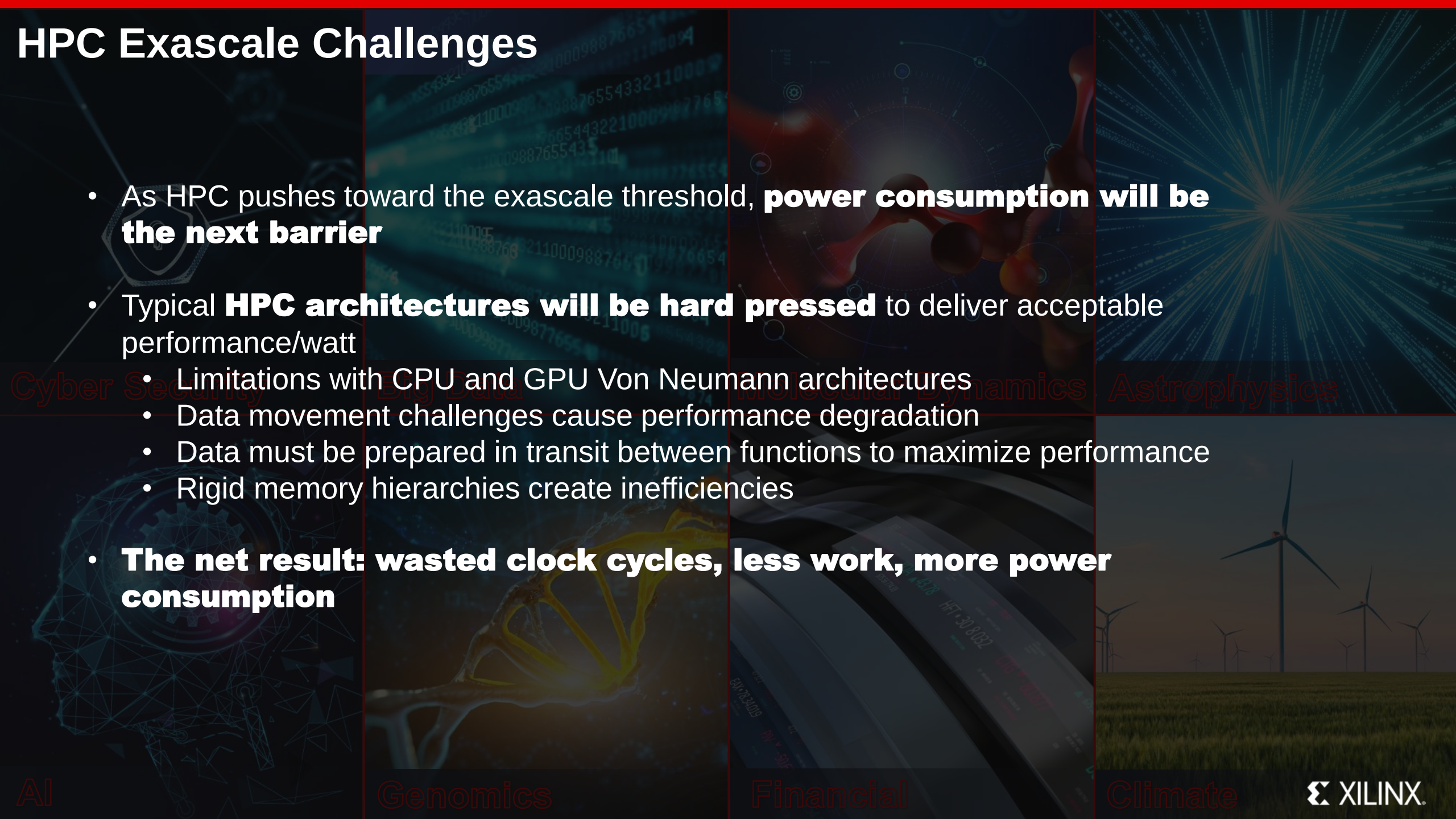
XILINX

Introducing the Alveo U55C Data Center Accelerator

Nathan Chang
HPC Product Manager, Data Center Group at Xilinx

HPC Exascale Challenges

- As HPC pushes toward the exascale threshold, **power consumption will be the next barrier**
- Typical **HPC architectures will be hard pressed** to deliver acceptable performance/watt
  - Limitations with CPU and GPU Von Neumann architectures
    - Data movement challenges cause performance degradation
    - Data must be prepared in transit between functions to maximize performance
    - Rigid memory hierarchies create inefficiencies
- **The net result: wasted clock cycles, less work, more power consumption**



A Breakthrough HPC Architecture
Today we are announcing:

- The most capable Alveo HPC accelerator card ever
- A groundbreaking HPC clustering solution that enables massive scale-out across existing customer infrastructure and network
- Full high-level programmability of both application and cluster

XILINX

Xilinx Scale Out System Architecture

Scale Out Architecture on RoCE v2 and DCBx with **existing data center server infrastructure**

- *Lower cost, no need for proprietary hardware, 1000+ node scale*
- *High performance, lossless, high bandwidth, low latency network connectivity*

Shared workload and shared memory **across multiple cards**

- *Bigger data, bigger workloads*

MPI enables hyperparallelism of Xilinx Adaptive Compute across nodes

- *Framework in widespread use with HPC developers today*

Vitis Unified Software Platform

Domain-specific development environment

Vitis accelerated libraries and APIs

Vitis core development kit

Customer built applications

C C++ python P4

Vitis AI

TensorFlow Caffe PyTorch

AI/ML

Video Streaming SDK

FFmpeg gstreamer

Video transcoding

Vision Library | Graph analytics Library | Fintech Library | Storage Library | Networking Library and middleware

Compilers
ARM, HLS, AI Engines, P4

Analyzers

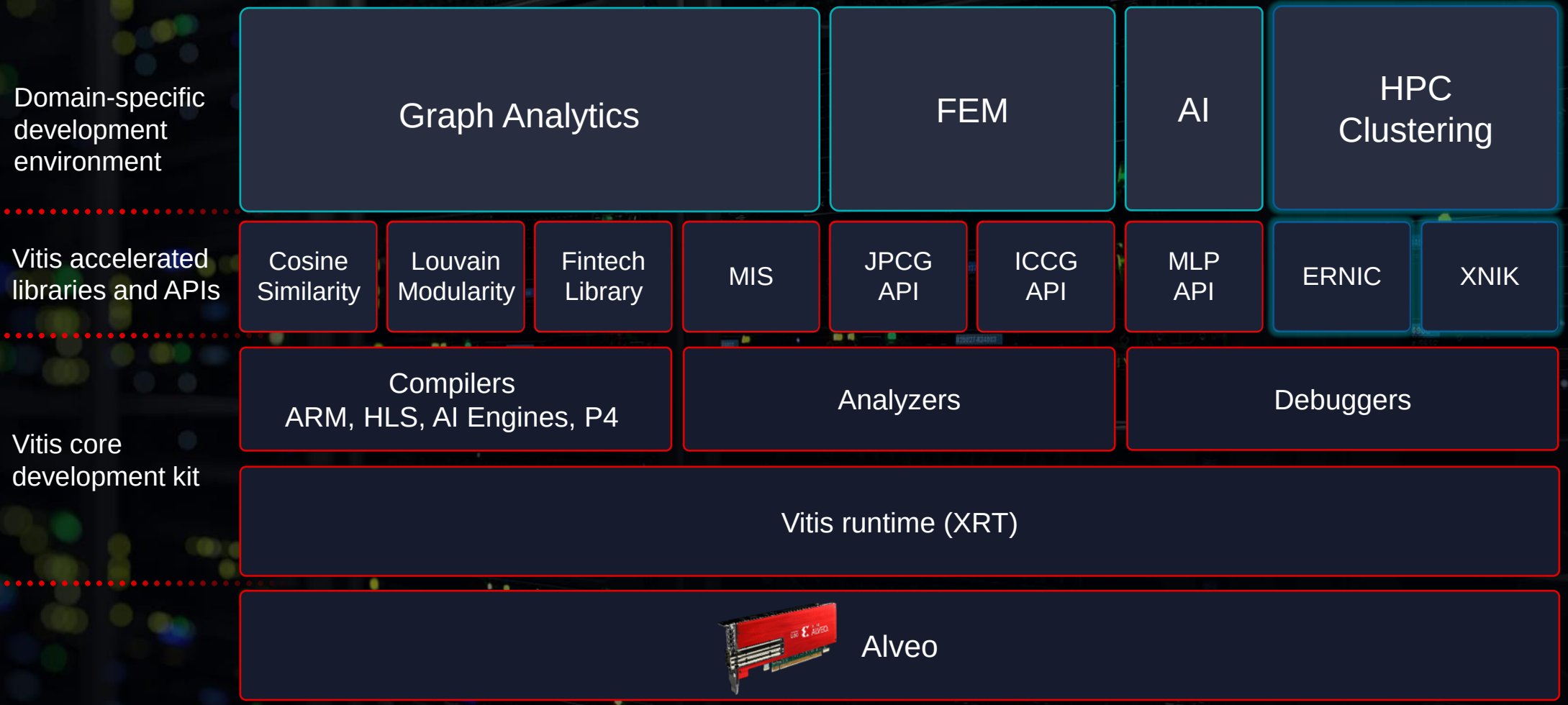
Debuggers

Vitis runtime (XRT)

Alveo

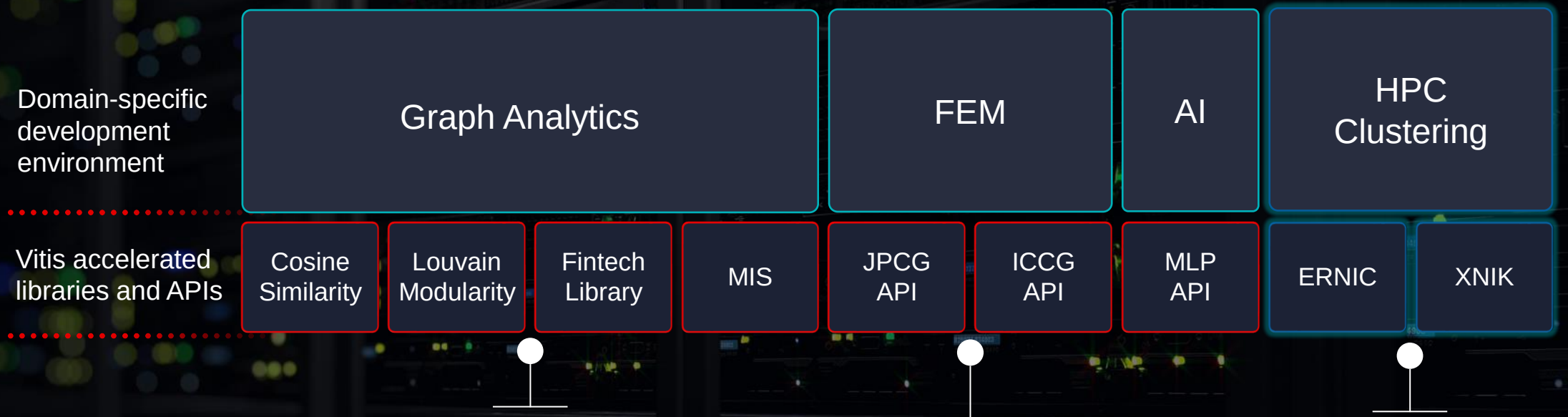
XILINX.

Vitis Unified Software Platform



XILINX.

Vitis Unified Software Platform



Enabling data scientists to create recommendation engines with low-touch/no-touch coding

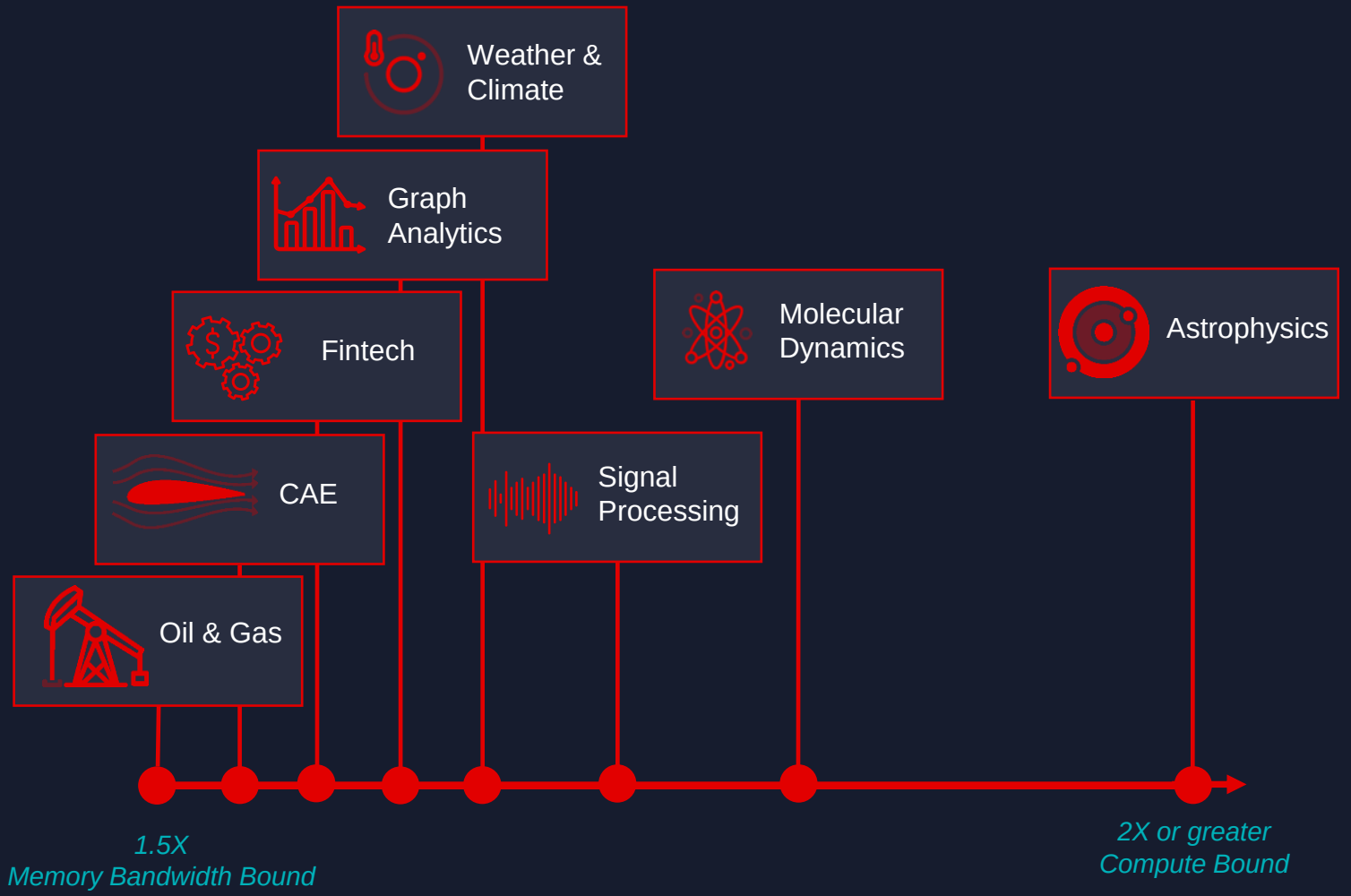
APIs for FEM developers to leverage custom data movement- no hierarchical memory lock-in, no cache misses

Enabling scale out across converged Ethernet

XILINX

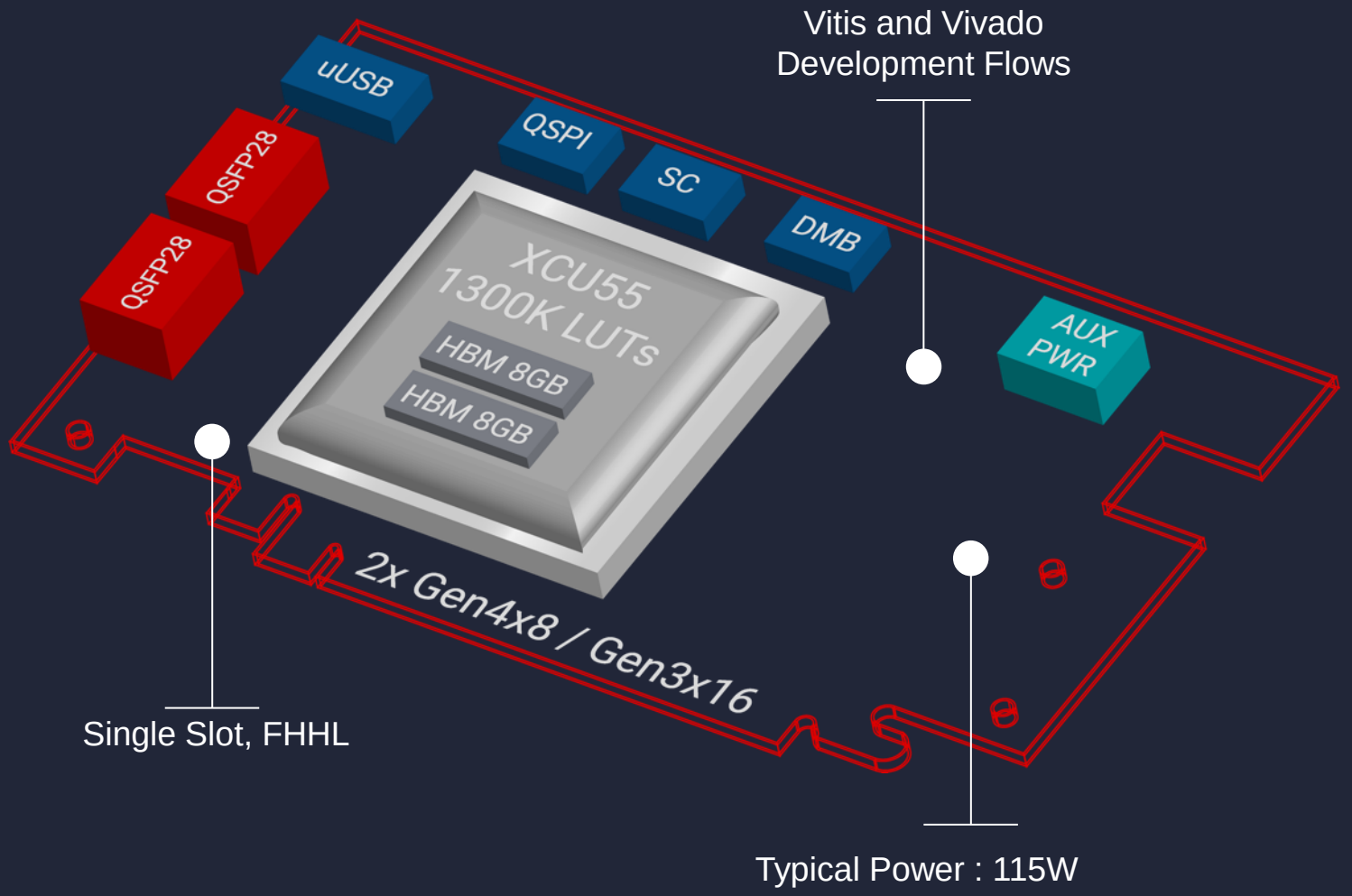
The Alveo U55C: Purpose Built for HPC and Big Data Workloads

- Many HPC workloads are either compute or memory bandwidth bound…
- I/O requirements expand exponentially over time…
- and Power consumption is a huge issue in the data center
- HPC needs gravity of compute and high-bandwidth memory
- In response, we built our most powerful accelerator ever and made sure it scaled easily



The Alveo U55C: Purpose Built for HPC and Big Data Workloads

- More parallelism of data pipelines
- Superior memory management
- Optimized data movement throughout the pipeline
- Best performance-per-watt



Our Most Capable Accelerator Ever

		Alveo U280	Alveo U55C
Dimensions	Width	Dual Slot	Single Slot
	Form Factor, Passive Form Factor, Active	Full Height, ¾ Length Full Height, Full Length	Full Height, Half Length
Logic	Look-Up Tables	1,304K	1,304K
	Registers	2,607K	2,607K
	DSP Slices	9,024	9,024
DRAM Memory	DDR Format	2x 16GB 72b DIMM DDR4	–
	DDR Total Capacity	32GB	–
	DDR Max Data Rate	2400MT/s	–
	DDR Total Bandwidth	38GB/s	–
	HBM2 Total Capacity	8GB	16GB
	HBM2 Total Bandwidth	460GB/s	460GB/s
Internal SRAM	Total Capacity	43MB	43MB
	Total Bandwidth	35TB/s	35TB/s
Interfaces	PCI Express®	Gen3 x16	Gen3 x16, 2x Gen4 x8
	Network Interface	2x QSFP28	2x QSFP28
Power and Thermal	Thermal Cooling	Passive, Active	Passive
	Typical Power	100W	115W
	Maximum Power	225W	150W

Xilinx Adaptive Computing For HPC

Clustering at massive scale

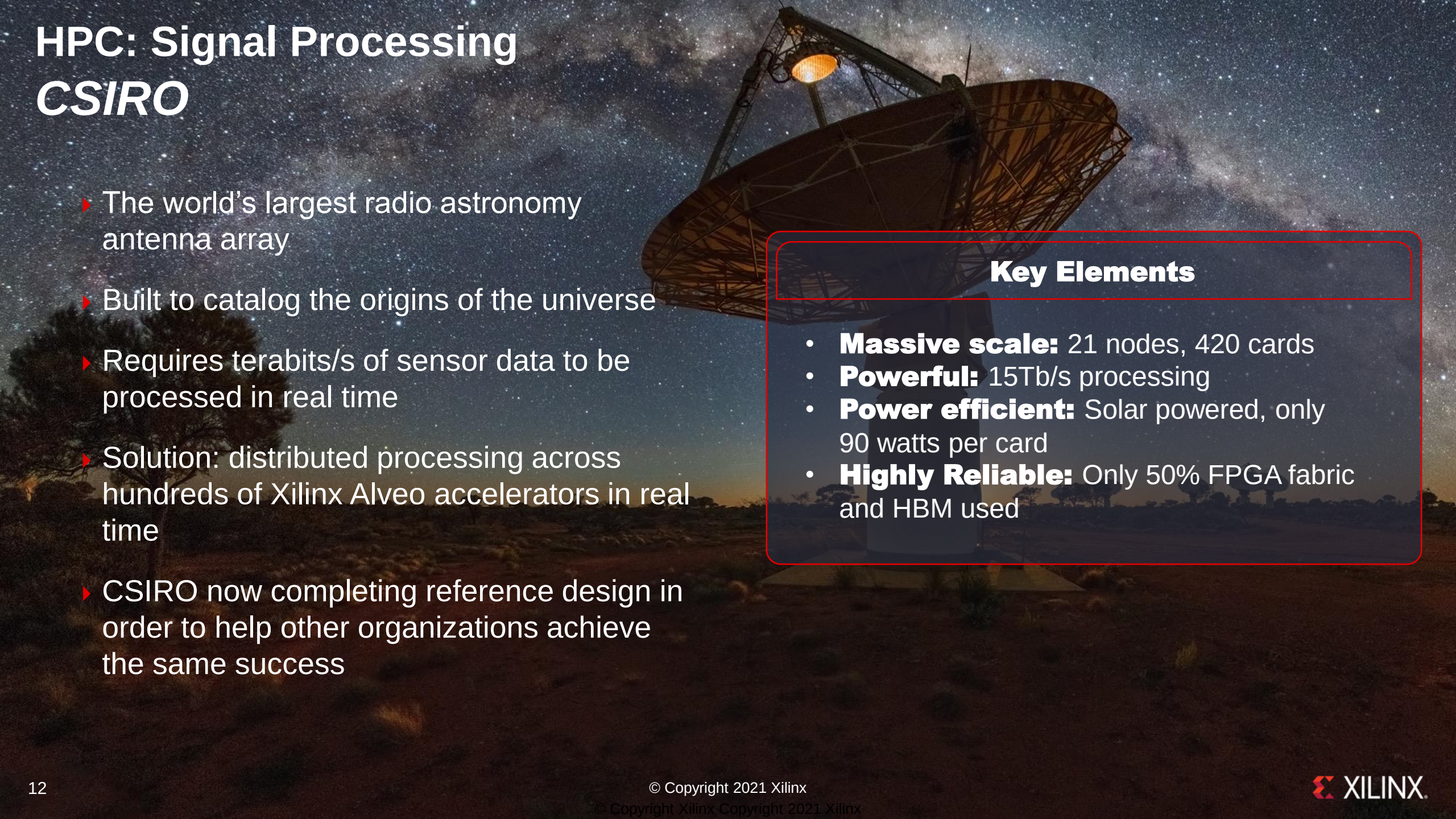
Built for software developers

Powerful accelerators

XILINX.

HPC: Signal Processing
CSIRO

- The world's largest radio astronomy antenna array
- Built to catalog the origins of the universe
- Requires terabits/s of sensor data to be processed in real time
- Solution: distributed processing across hundreds of Xilinx Alveo accelerators in real time
- CSIRO now completing reference design in order to help other organizations achieve the same success

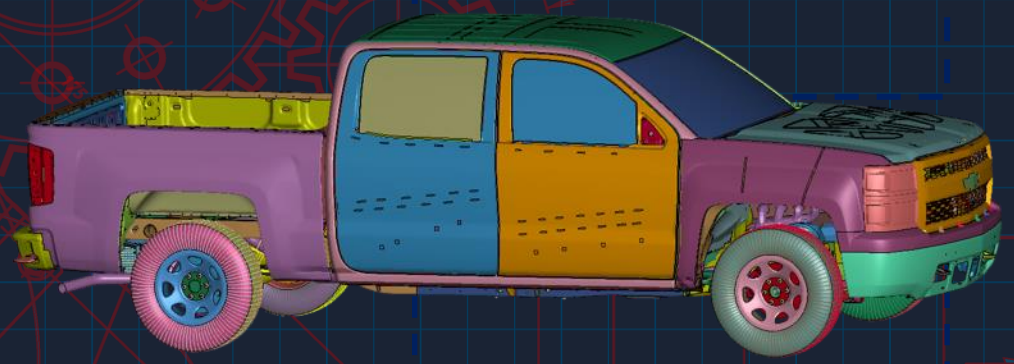


Key Elements

- **Massive scale:** 21 nodes, 420 cards
- **Powerful:** 15Tb/s processing
- **Power efficient:** Solar powered, only 90 watts per card
- **Highly Reliable:** Only 50% FPGA fabric and HBM used

HPC: Computer-Aided Engineering(CAE)
ANSYS LS-DYNA

- LS-DYNA: Finite Element Program (FEM)
  - Uses FEM to simulate real-world product performance
  - LS-DYNA allows designers and engineers the ability to create simulations with an infinite amount of complexity
- Large scale simulations take weeks on a CPU
  - x86 architectures aren't equipped to provide the high I/O & bandwidth required
  - CPU memory hierarchies are inflexible and that creates unnecessary overhead
  - X86 architectures are inherently inefficient at handling data movement



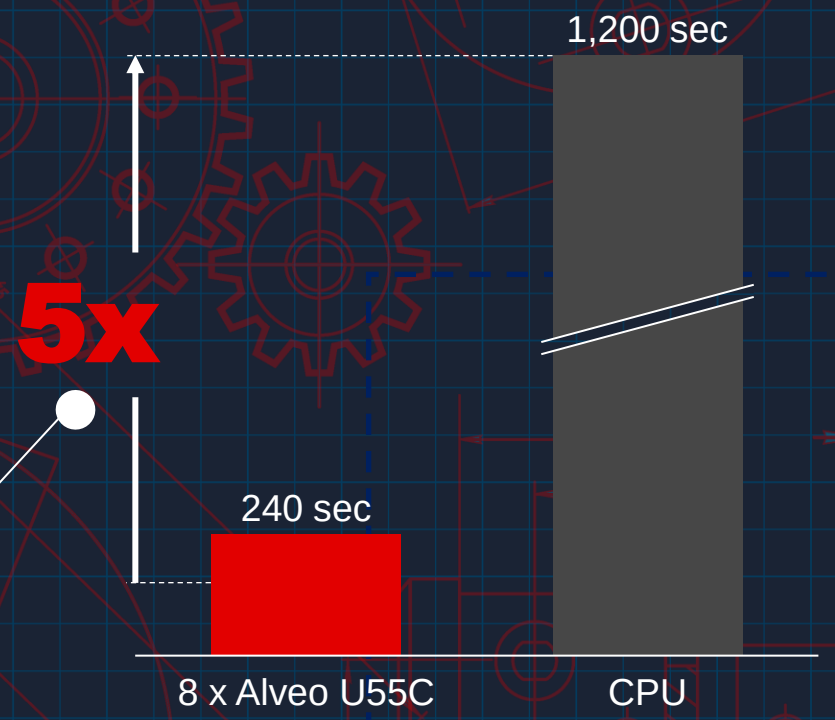
- Silverado model from the National Crash Analysis Center
- **700k elements** (mostly shells, plus solids and beams)
- Gravity loading

U55C Hyperparallel Data Pipelining for LS-DYNA

- Data is pipelined to simply stream between functions
- Data is prepared in transit to achieve maximum throughput
- Highly composable memory hierarchies
  - 16GB HBM2 memory, 32 HBM channels @ 460GB/s

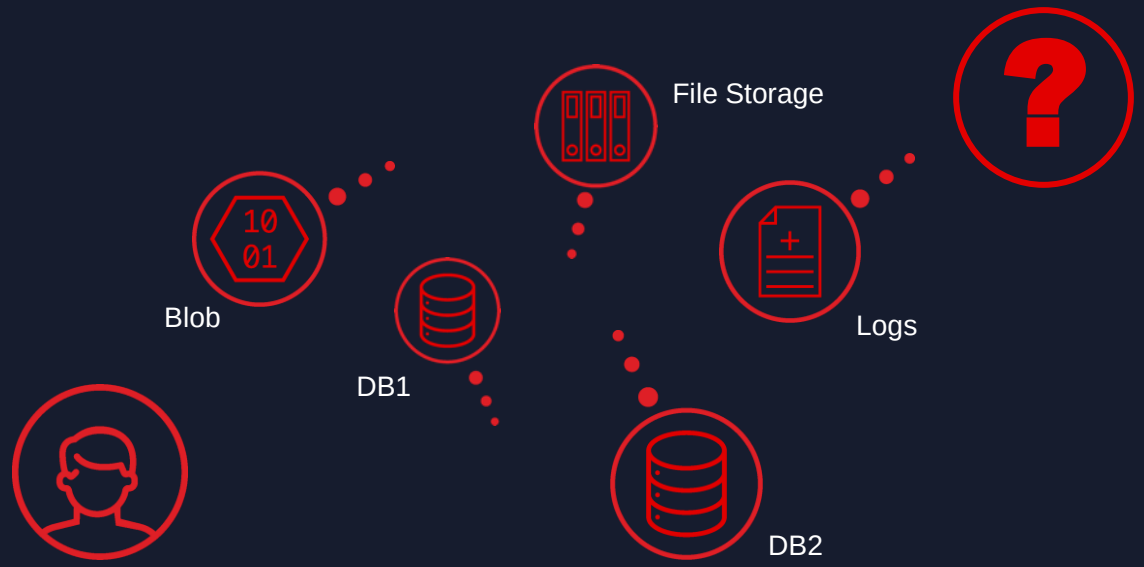
Workload partitioned across multiple Alveo U55C cards

Result: 5x Performance vs CPU



***D**imensions of matrix -> 12M*
***nnzs:** No of non-zero elements -900M*
***Time in secs**: JPCG solver equation runtime*
***CPU model:** Intel Xeon Platinum 8260L @2.4GHz, 1.5TB memory*

Big Data: Graph Analytics



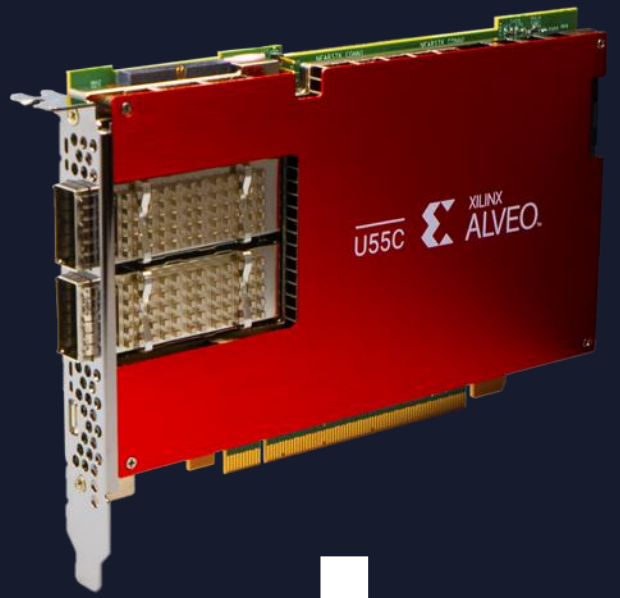
- Tabular and unstructured databases do not not focus on relationships
- Data correlation is key to understand behavior and unlock prediction
- It's expensive to have data scientists search for answers in disconnected data environments

- Graph makes data relationships the focus, quickly delivering insights that were previously expensive and difficult to obtain
- **The next frontier for graph is finding those answers IN REAL TIME**

U55C Real-Time Graph Insights

- Real time results demand Xilinx acceleration
- Tuned algorithms for important real-time graph use cases

- Patient Recommendation
  - Cosine similarity algorithm
  - **96x over CPU**
  - 2x better over U50 in overall perf
- Fraud Detection/Anti-money laundering
  - Louvain modularity algorithm
  - **45x over CPU**
  - 66% reduction in DDR4 utilization
  - 35% higher quality score



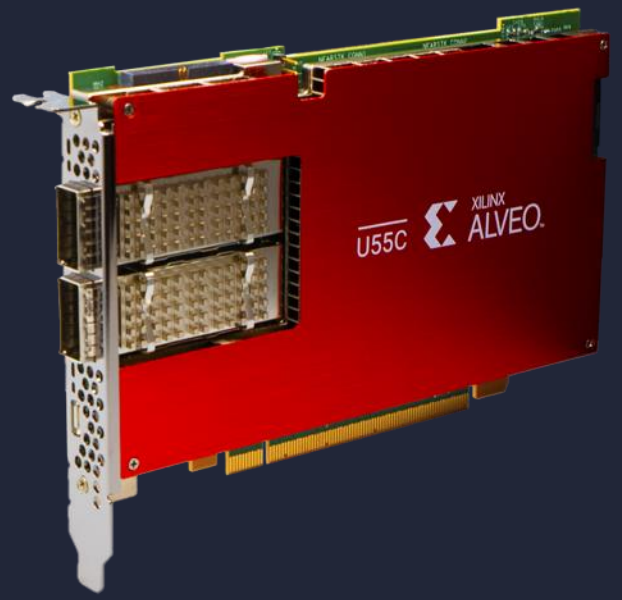
+

TigerGraph

XILINX

How To Try and Buy

U55C Available Now



Available on Xilinx.com and through Xilinx distributors

Easy Evaluation In Cloud and In DC

Cloud FPGAaaS

- Xilinx App Store access
- Fixed server configurations
- Vitis design flow
- Managed infrastructure

Colocation

- Available for partner and customer evaluation
- Specific server configurations
- Vivado design flow
- Managed infrastructure

Summary

- The new Xilinx HPC clustering solution enables **massive scale-out** across **existing customer infrastructure and network**
- The **Xilinx Alveo U55C accelerator** card, now shipping, brings **superior performance per watt to HPC and database workloads** and easily scales through Xilinx clustering
- Software developers and data scientists can unlock the benefits of Xilinx adaptive computing through **high-level programmability of both application and cluster**

Thank You