HOW WE GOT STARTED…

DAN GOLDIN

KnuEdge FOUNDER & CEO

- Directly served 3 Presidents as Director of NASA
- Led globally significant technology innovations
- Previously VP and GM of TRW Space and Tech Group

PLATFORM AND TECHNOLOGY

Robert Patti

VP Hardware

INFORMATION IS BEING REVOLUTIONIZED

Information is becoming more and more valuable through increasingly intelligent machines

SUCCINCT

PERSONALIZED

FRICTIONLESS

EVERYWHERE

CRITICAL BARRIER IN TIME

Despite all the extraordinary effort and enthusiasm, the progress of A.I. is profoundly hindered by developers' pain in the time and concomitant uncertainties of training neural networks

> *If hyperparameters are poorly chosen, the network may learn* ***slowly, or perhaps not at all***
>
> **Deeplearning4j**

> *I'd like to know ahead of time if my training will take* ***8 hours, 8 days or 8 weeks***
>
> **StackExchange**

8 hours

to train **simple** A.I. algorithm

4 weeks

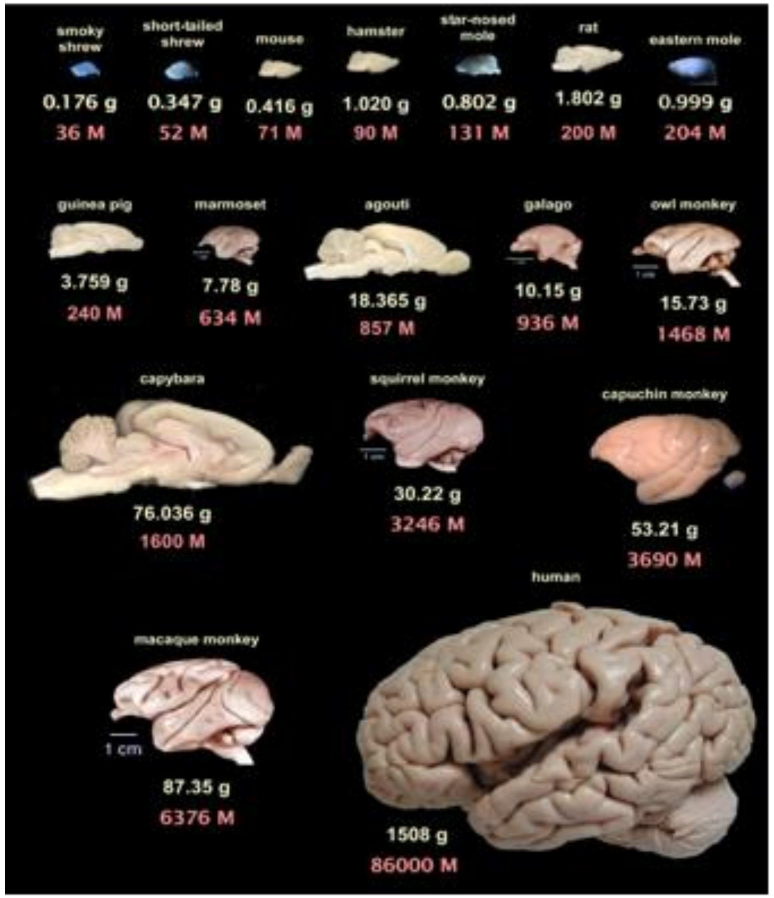
to train **complex** A.I. algorithm

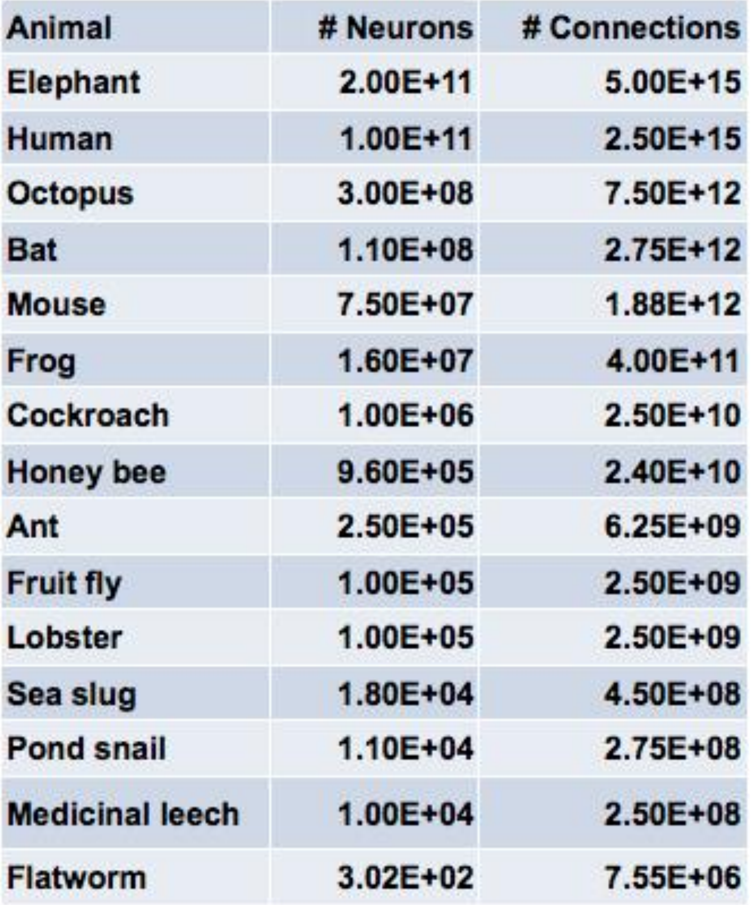
2 months

to train **very complex** A.I. algorithm

THE NEUROBIOLOGICAL COMPUTER

Neurobiological Systems: Flexible and Scalable





Animal	# Neurons	# Connections
Elephant	2.00E+11	5.00E+15
Human	1.00E+11	2.50E+15
Octopus	3.00E+08	7.50E+12
Bat	1.10E+08	2.75E+12
Mouse	7.50E+07	1.88E+12
Frog	1.60E+07	4.00E+11
Cockroach	1.00E+06	2.50E+10
Honey bee	9.60E+05	2.40E+10
Ant	2.50E+05	6.25E+09
Fruit fly	1.00E+05	2.50E+09
Lobster	1.00E+05	2.50E+09
Sea slug	1.80E+04	4.50E+08
Pond snail	1.10E+04	2.75E+08
Medicinal leech	1.00E+04	2.50E+08
Flatworm	3.02E+02	7.55E+06

Where KnuEdge wants to be in 2021: MindScale.

Current generation machine learning capabilities

x

NEURAL COMPUTING AS A SERVICE

DESIGN GOALS

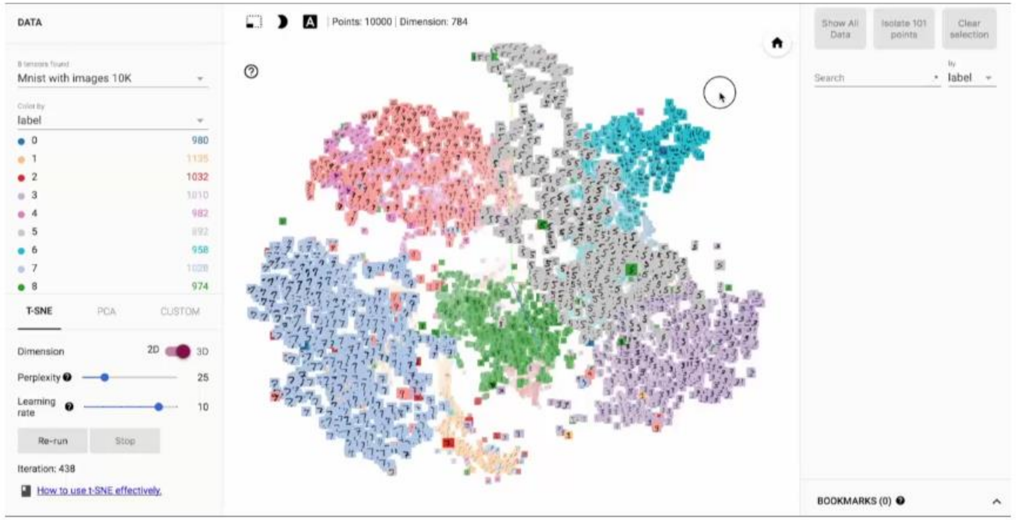
- Make it simple & fun
- Make it powerful
- Make it affordable
- Make it accessible
- Make it secure & reliable

SOFTWARE TOOLS

- TensorFlow
- Deeplearning4j
- Julia
- Compiler implemented as LLVM backend
  - Anything that can compile to bitcode can be emitted as Knureon code
  - Also includes assembler
  - Debugger working on LLDB
  - Linkable bitcode libraries allow global optimization in tDSP kernels
- KNI
- KPI
- BLAS libraries in development

TENSORFLOW

- Open source data flow platform
- Focused on Machine learning, expandable to many other applications
- Python and C libraries supported
- Includes visualization and analysis framework which requires no porting from us
- Allows users to port apps to Knureon hardware without any knowledge of our architecture.



DATA

8 tensors found: Mnist with images 10K

Color By: label

● 0	980
● 1	1135
● 2	1032
● 3	1010
● 4	982
● 5	892
● 6	958
● 7	1028
● 8	974

T-SNE

Dimension: 2D 3D

Perplexity: 25

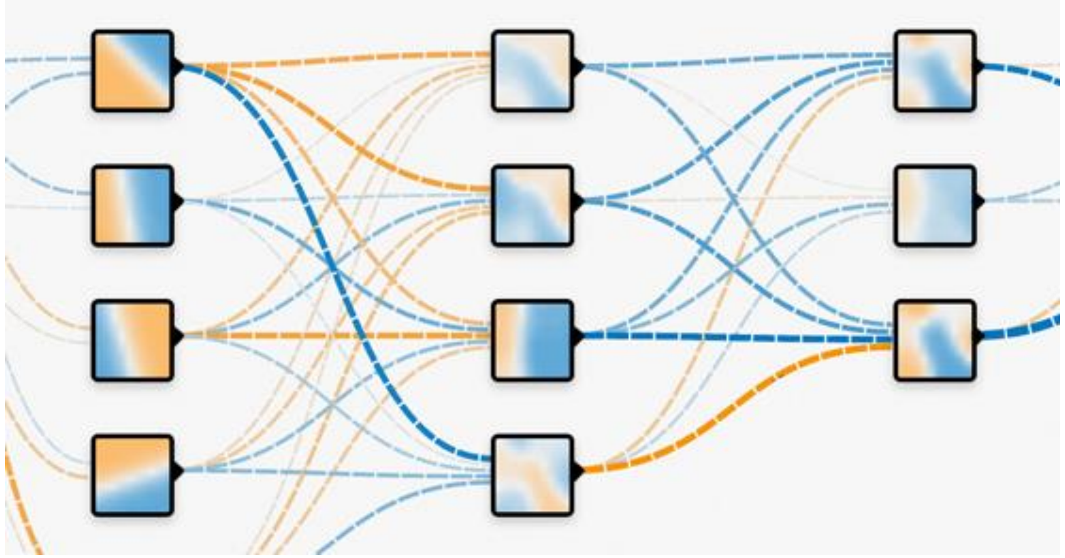
Learning rate: 10

Iteration: 438

How to use t-SNE effectively.

Points: 10000 | Dimension: 784

Search By: label

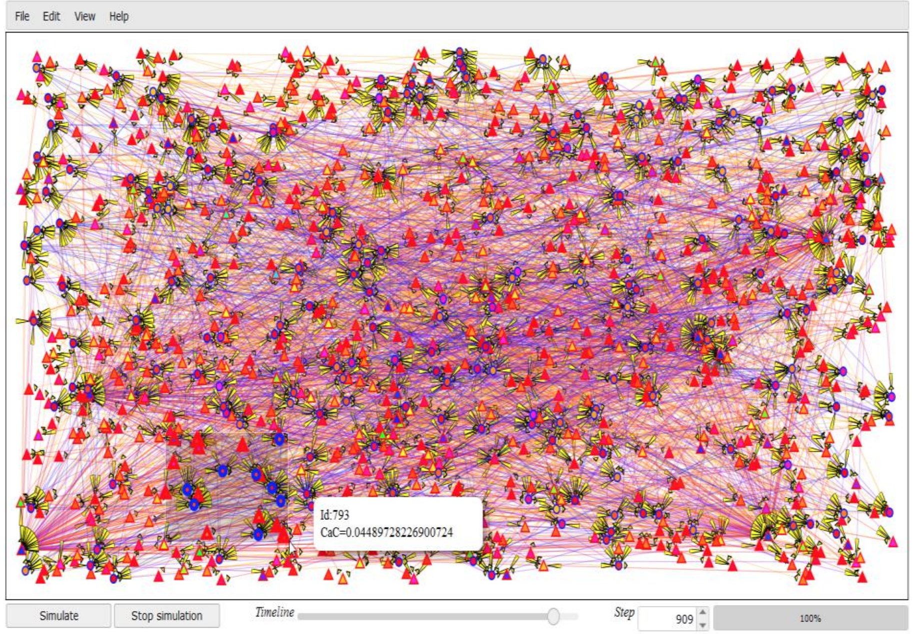


DEEPLEARNING4J / SKYMIND

- Open-source distributed deep-learning platform
- Focused on Deep Learning, can be used for many other applications
- Java, Scala and Clojure supported
- Enterprise support from the people at Skymind, a San Francisco-based business intelligence and enterprise software firm
- Allows users to port apps to Knureon hardware without any knowledge of our architecture

JULIA COMPUTING

- Open-source language for data science, machine learning and scientific computing
- Supports R, Python, Matlab, SAS or Stata
- Speed, capacity and performance of C, C++ or Java
- Over 1m downloads
- Allows users to port apps to Knureon hardware without any knowledge of our architecture



Id:793
CaC=0.04489728226900724

THE LIMITS OF COMPUTING

- Memory technology is stagnant
  - DRAM is still 60-70ns tRC
- Processor speed is stalled
  - It’s the wire
- End of semiconductor scaling
- Dominated by power usage

THE SPECIAL SAUCE IN THE HARDWARE

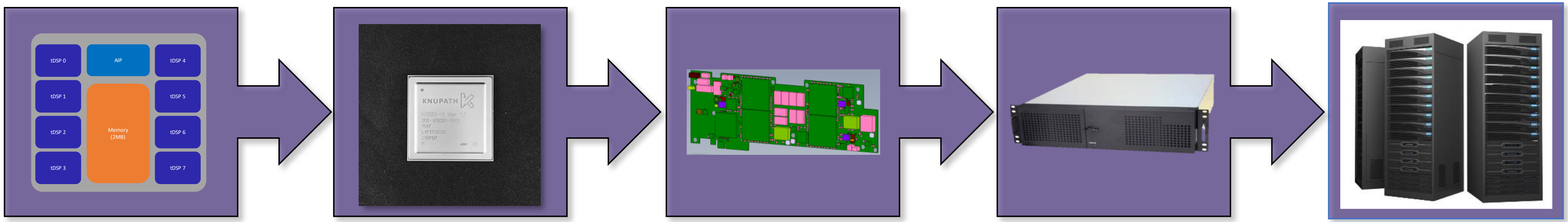
- It is all about 2 system objectives:
  - Reduced Latency
    - Reducing both access time to memory and between processing elements
  - Increased Scalability
    - The ability to expand the system almost without limit while having little impact on the local or global task performance
- How we do it:
  - Unique and abundant light weight processors embedded in memory
  - Flat machine design and addressing space – Memory rich architecture
  - Simplified memory backbone communications – LambdaFabric using FLITs
  - 2.5/3D integration

HARDWARE ADVANCES

- Multi-Chip Packaging (Chiplets)
    - Rapid (3 months) turnaround
- Interconnections
    - 2.5D and 3D integration
- Low latency
    - Memory
    - Interconnect
- Heterogenous Engines
    - tDSP
    - Fine scale
    - Massively parallel
- Resilience / Fault Tolerance

LambdaFabric® SCALABLE COMPUTING

LOW-LATENCY, HIGH-THROUGHPUT, LOW-POWER COMPUTING FABRIC



Cluster — 6ns latency

Chip — 62ns

Board — 247ns

Rack

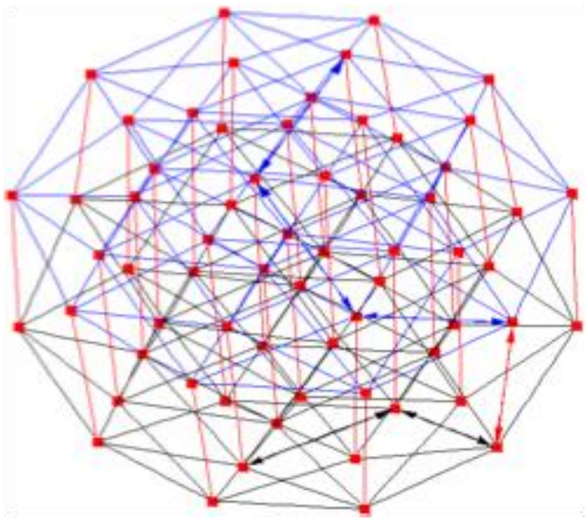
Multi-Rack — 437ns

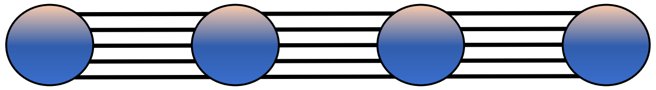
- ✓ Scale invariant network architecture
- ✓ Low latency to everywhere
- ✓ High bandwidth
- ✓ Multi-dimensional connectivity
- ✓ Scalable up to 524,288 chips

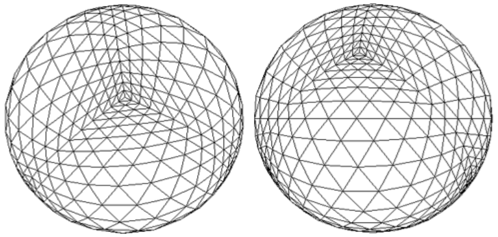
FLEXIBLE COMMUNICATION TOPOLOGY

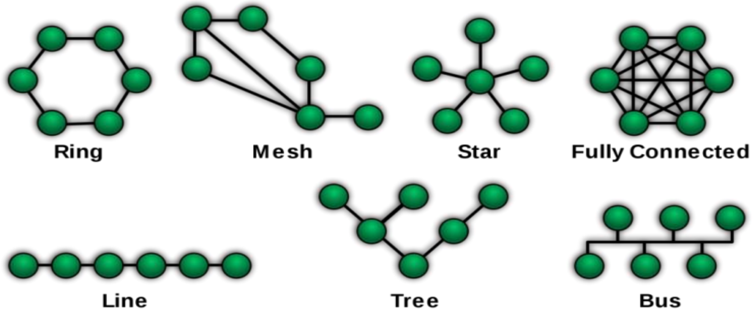
Linked Hermosa chips create a LambdaFabric®

- LambdaFabric® network is of arbitrary topology
- ARTs in the routers direct packets
- Guaranteed C2C delivery
- Multi-dimensional grids eliminate hops
- Automatic topology discovery
- Up to 524,288 devices in a block





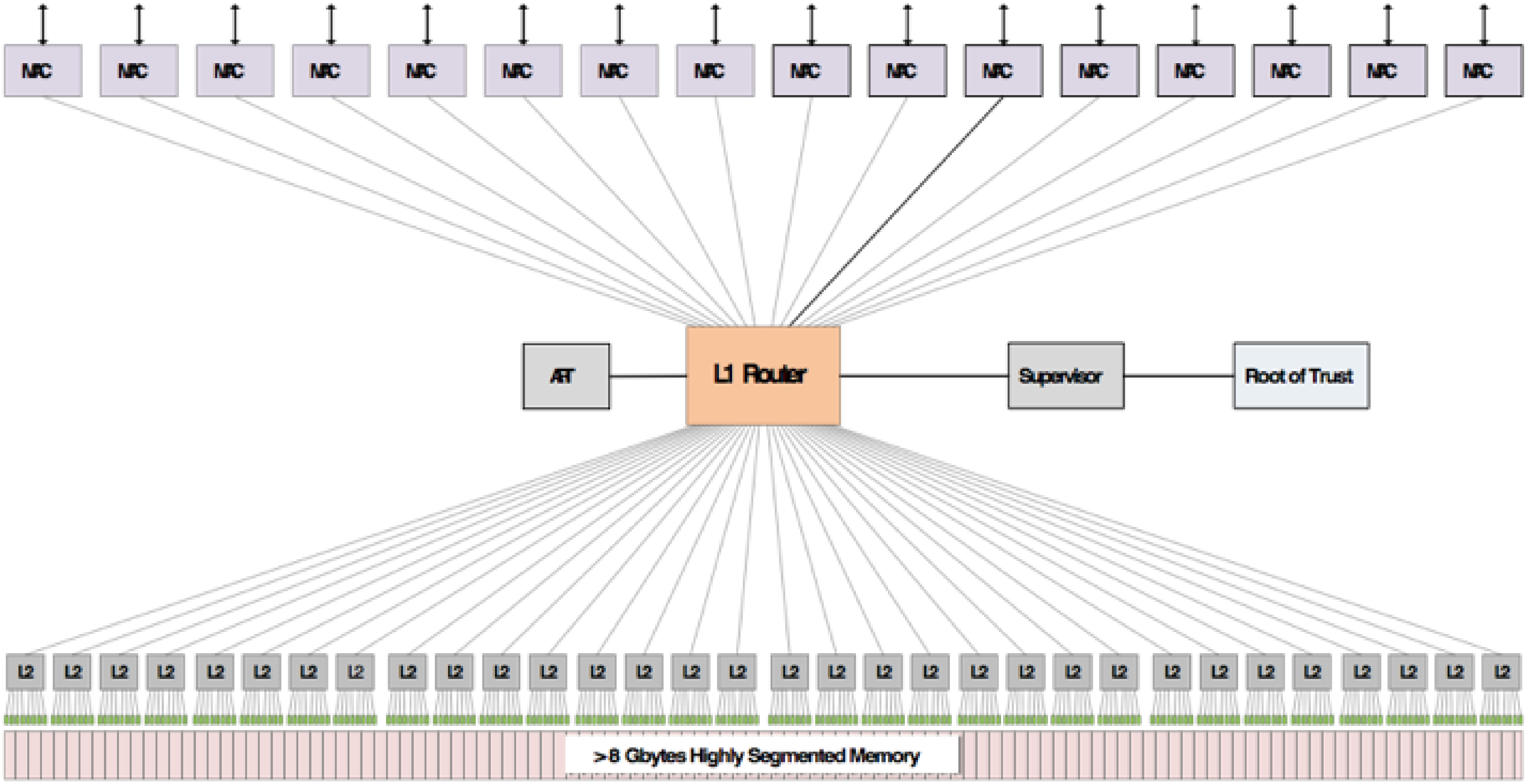




HERMOSA CPU (GEN1)

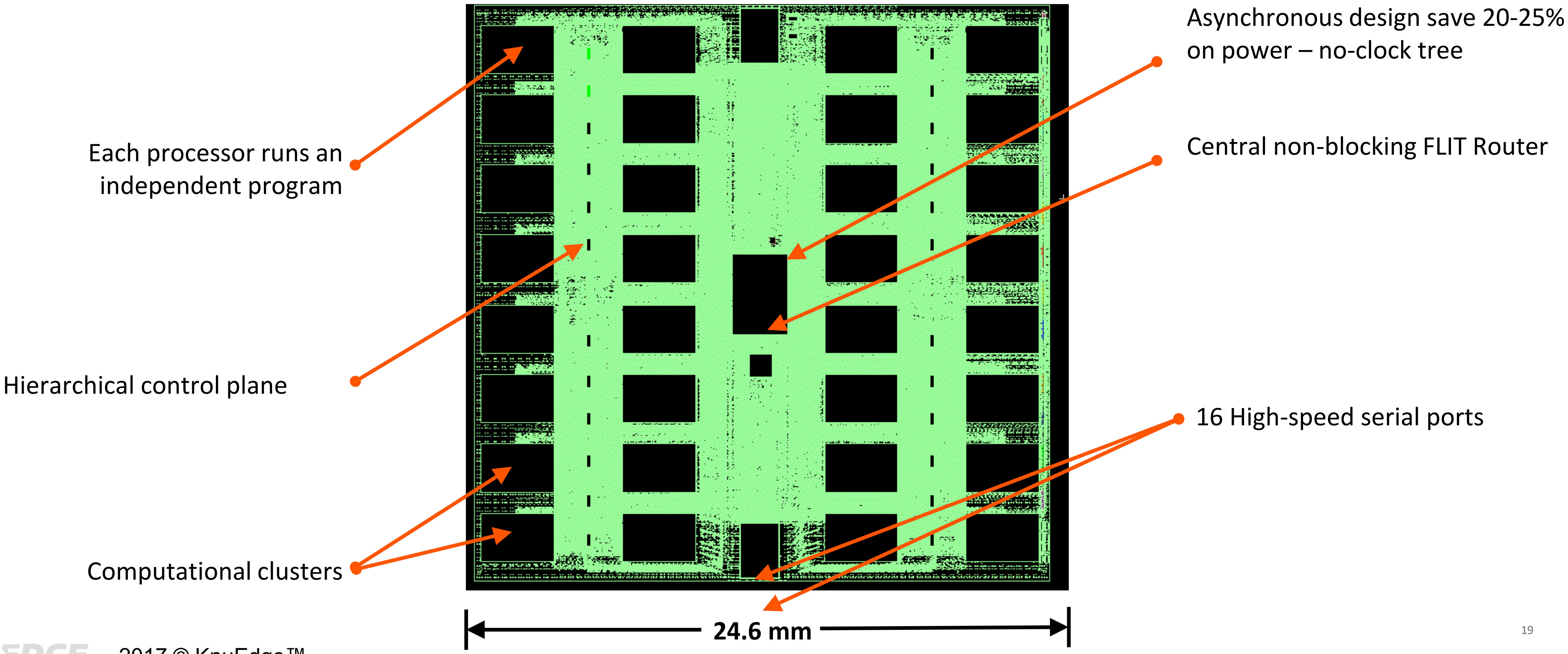
- 256 cores each with 288KBytes of near memory (L1)
- 256 registers for in-register computing
- Each core has 256 x 4 bytes multiported register file
- 1GHz clock rate
- 16 10Gb/s SERDES channels
- One 32b FPU
- One 32b transcendental math engine

HERMOSA - A DATA CENTER ON A CHIP



HERMOSA IMPLEMENTATION

CLOUD-ON-A-CHIP ARCHITECTURE – ROUTER CENTRIC



Each processor runs an independent program

Hierarchical control plane

Computational clusters

Asynchronous design save 20-25% on power – no-clock tree

Central non-blocking FLIT Router

16 High-speed serial ports

24.6 mm

HYDRA CPU

- Gen1 Hermosa processors
- Repackaged into 8 die modules
- Improves density
- Allows near matching to GPU floating point capability at individual board level
- 2048 cores each with 288KBytes of near memory (L1)

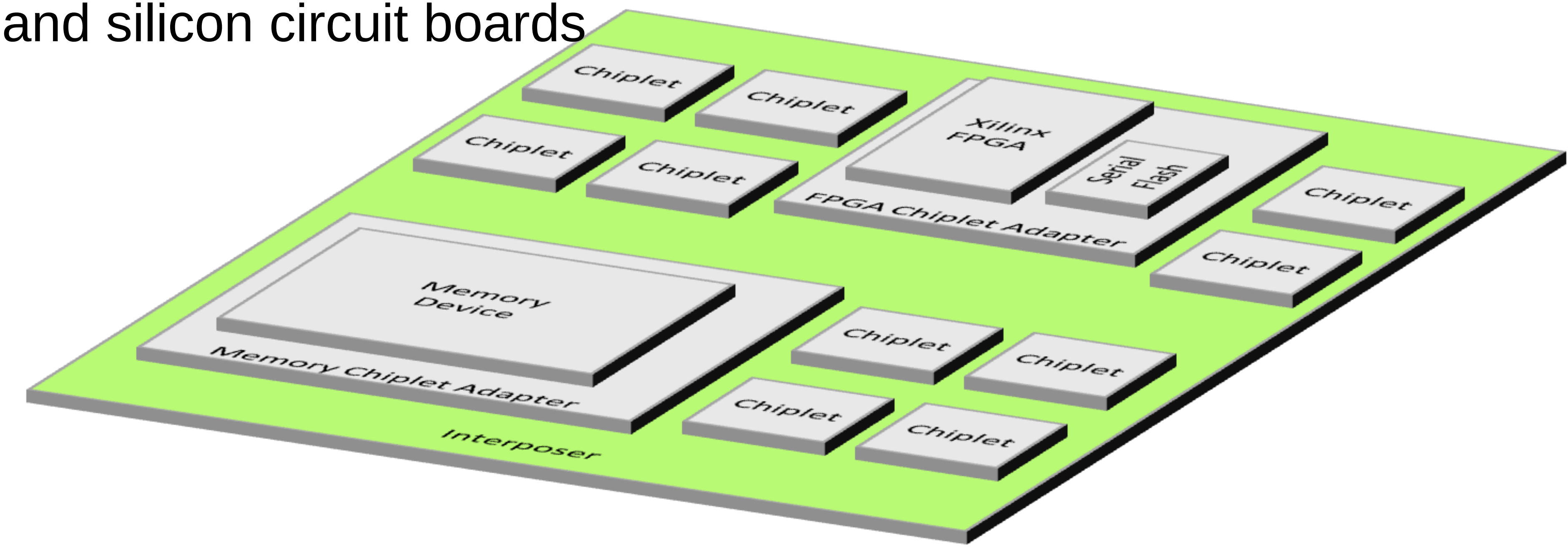
HYDRA-X CPU

- Builds directly on Hydra module
- Uses water cooling for 2x more card level density
- 15% clock speed uplift

CHIPLET SYSTEMS

Mixed technology assemblies

- Flip-chip
- Copper pillar
- DBI die to wafer
- Organics and silicon circuit boards



CAPITOLA CPU CHIPLET - GEN2

- 256 cores each with 128MBytes of near memory (L1)
- Each core has 4 x 256 x 8 bytes multiported register file
- ≥2GHz clock rate
- Four 64b Vector FPUs per core
  - 4 64bit FLOPs
  - 8 32bit FLOPs
  - 16 16bit FLOPs
- One 32b transcendental math engine
- New local stack
- Native 25Gb/s FLIT Link interfaces (>10x improvement on GUPS messaging vs Hermosa)
- Chiplet construction

CAPITOLA MODULES

- Uses chiplet assembly
- Up to 8 processing units selecting from
  - Capitola CPUs
  - Analog LPUs -Low precision with extremely low energy
  - RRAM Vector Processors High efficiency compute
  - GPU Gen7/8 Large scale SIMD 32/64b FPUs
- Targeted performance
- Easily adapted to new market needs
- 4 x 32 28Gb/s SERDES chiplets
- 128GB near memory
- ~100W modules

THE KNUEDGE CAPITOLA ADVANTAGE:

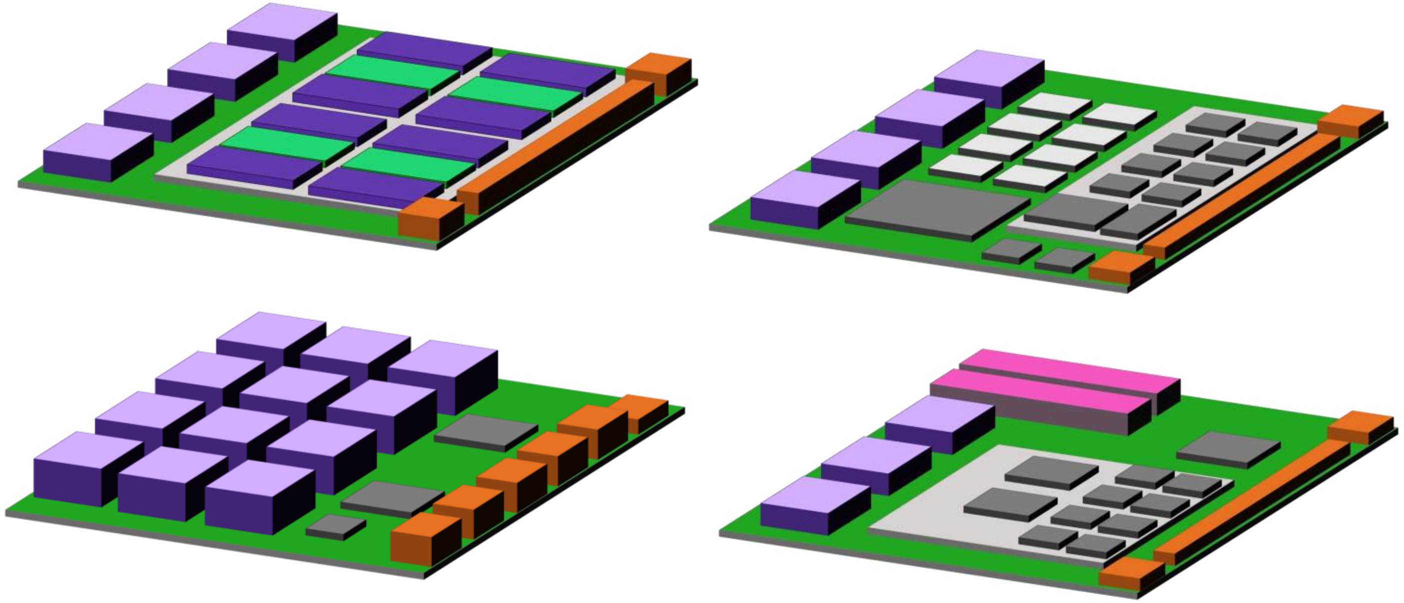
PERFORMANCE, COST, POWER, FLEXIBILITY

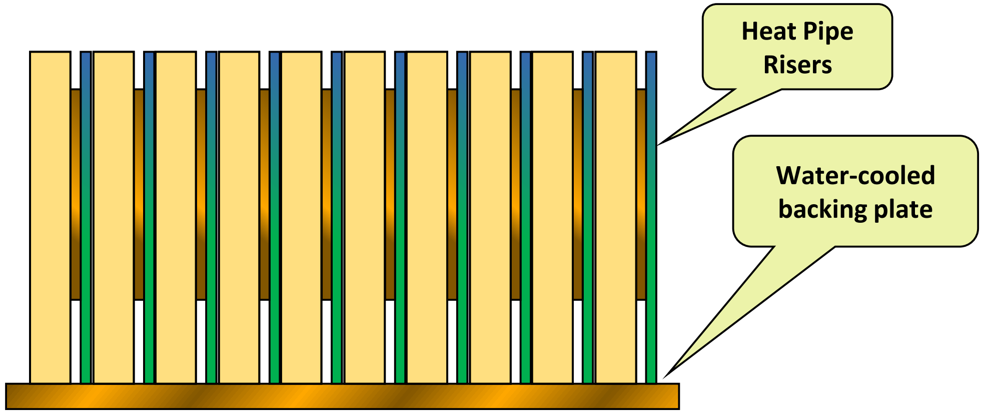
- ✓ Shortening distances that signals must travel between the processor, memories and I/Os
- ✓ The ability to optimize performance, power and cost more easily through different packaging options
- ✓ Improved time to market by using customized configurations from parts developed at a variety of process geometries.
- ✓ Increased yield from smaller parts

Potential Chiplets

- ADC/DAC for signal processing
- GDDR5 memory interface
- PCIe Gen 4
- FPGA bare die
- Extreme high-speed serial I/Fs
- Video compression/decompression
- Re-timers for long-range serial
- Arithmetic units (double precision)
- Bare flit router chip
- String search
- High-Speed Serial (HSS)

CAPITOLA MODULE DESIGNS





LambdaFabric® MicroBlades

- Plug in compute power
- About 70mm x 100mm x 12mm
- Heat pipe/bi-phase conductors
- Water cooled header
- Packs <u>10</u> racks into the space of one
- Prevents IP theft
- Not dependent on any single fab
- We can mix and match components
  - Can utilize the best technology for each component
  - Harness domain expertise for each area
- Reduces NRE, risk, and per-unit-costs

MIWOK PLATFORM

- Uses the Gen1 Hermosa die
- 65,536 cores in one air cooled rack
- Baseline first deployment for developers
- Uses off-the-shelf x86 server boxes
- Existing software stacks
- On-line in July

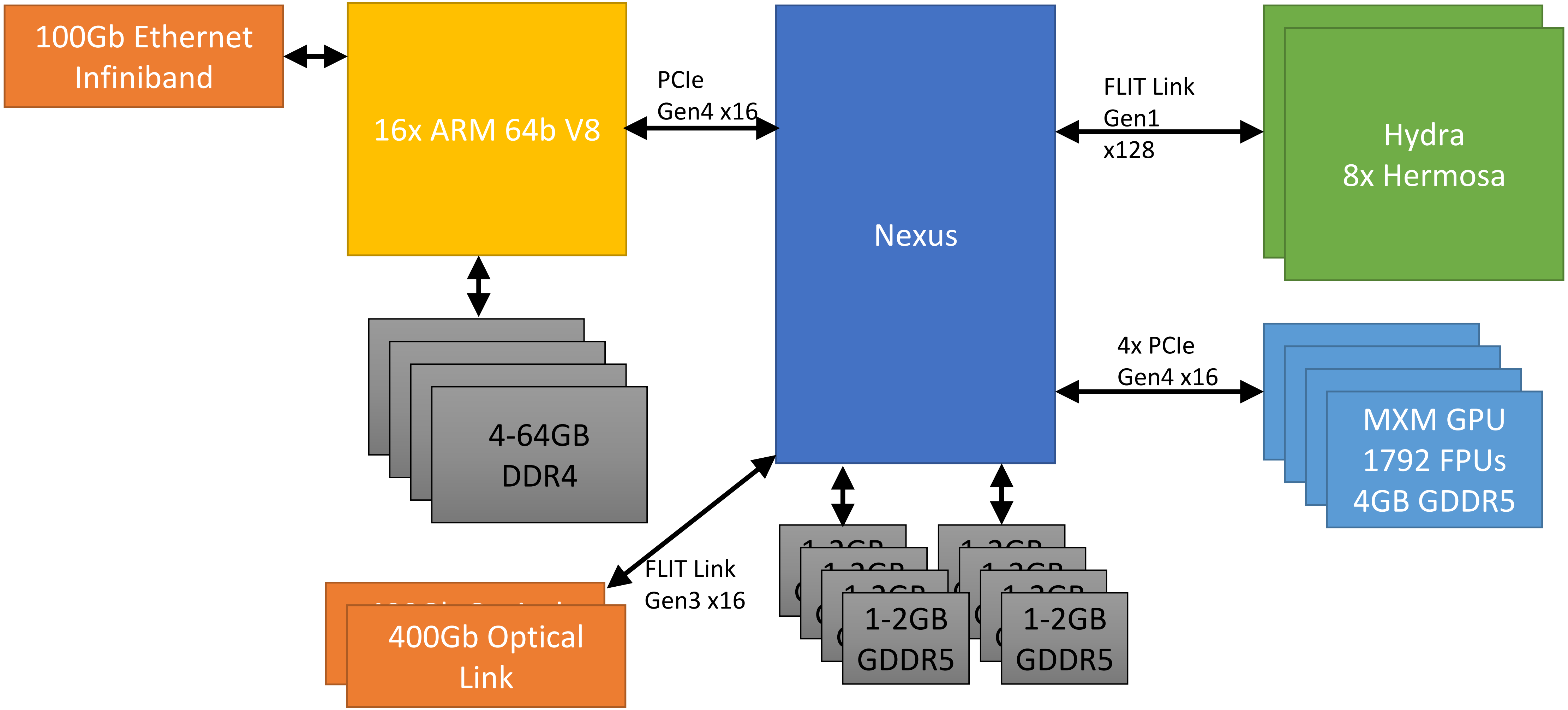
LAMBDABOX1 PLATFORM

- Uses the Gen1 Hermosa die
- 128k cores in one air cooled rack
- 1M core system planned for March 31, 2018 customer availability
- Includes 512 ARM V8.1 64b processors for general purpose processing and dispatch per rack
- Improves routing/switching capability and increases local memory bandwidth by 4x
- New FPGA design but no new silicon development

LAMBDABOX2 PLATFORM

- Uses the Hydra processor module
- 256k cores in one air cooled rack
- 3M core system upgrade planned for July 31, 2018 customer availability
- Includes 1024 ARM V8.1 64b processors for general purpose processing and dispatch per rack
- New NEXUS communications chip provides further improved routing/switching capability reducing routing latency by ~50%, increases local memory bandwidth by an additional 2x and reduces bulk memory latency by 75%
- NEXUS chip also provides interface for up to 4 GPU modules
- Major achievements:
    - True heterogenous computing
    - >4x nVidia P100 card performance on CNN
    - >50x x86 performance on RNN

LAMBDABOX2 25TFLOPS/4200 CORES



LAMBDABOX3 PLATFORM

- Uses the Hydra-X processor module
- 512k cores in one water cooled rack
- 4M core system upgrade planned for Sept 30, 2018 customer availability
- Includes 1024 ARM V8.1 64b processors for general purpose processing and dispatch per rack
- Reuses NEXUS communication chip
    - Larger GDDR memory size for denser problem computation
- GPU selection upgrades with higher performance (>40TFLOPs per card)
- Major achievements:
    - High power density liquid cooling wring-out
    - >6x nVidia P100 card performance on CNN
    - >110x x86 performance on RNN

LAMBDABOX4 PLATFORM

- Uses the Capitola processor module
- 1M cores in one water cooled rack
- +10M core system planned for March 31, 2019 customer availability
- Includes 2048 ARM V10 64b processors for general purpose processing and dispatch per rack
- Reuses NEXUS communication chip
  - Uses raw packet mode for 10x reduction in system latency
- Multiple processing accelerator options to address evolving markets
  - >100TFLOPs per card
- Major achievements:
  - >10x nVidia P100 card performance on CNN
  - >1000x x86 performance on RNN
  - ExaScale computing capable
  - Cuts cost by 5x per core, 25x per operation vs Hermosa Gen1

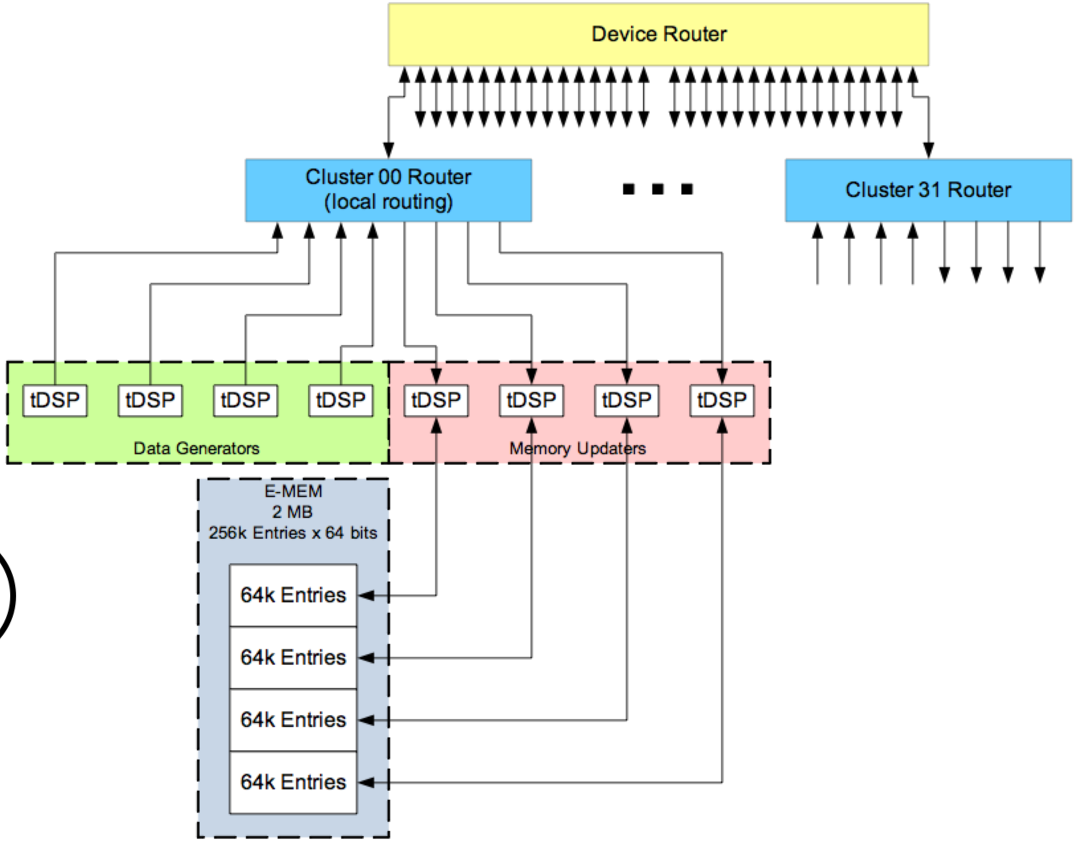
EMPHASES IN COMPUTATION

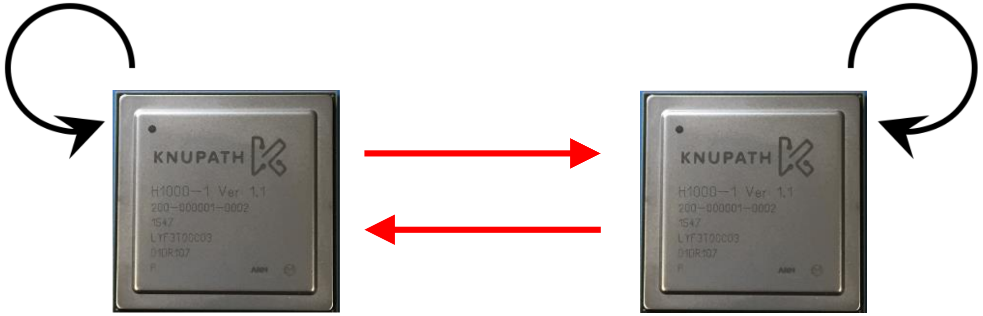
	FOCUS	COMPUTATION
1	Communication centered	GUPS (Random Memory Access)
2	Dense computation centered	CNN (Convolutional Neural Networks)
3	Balanced	K-Means Clustering / Image Segmentation
4	Scatter / Gather	Graph Processing
5	Sparse computation centered	Heterogeneous Sparse Matrix Neural Network

RANDOM MEMORY ACCESS

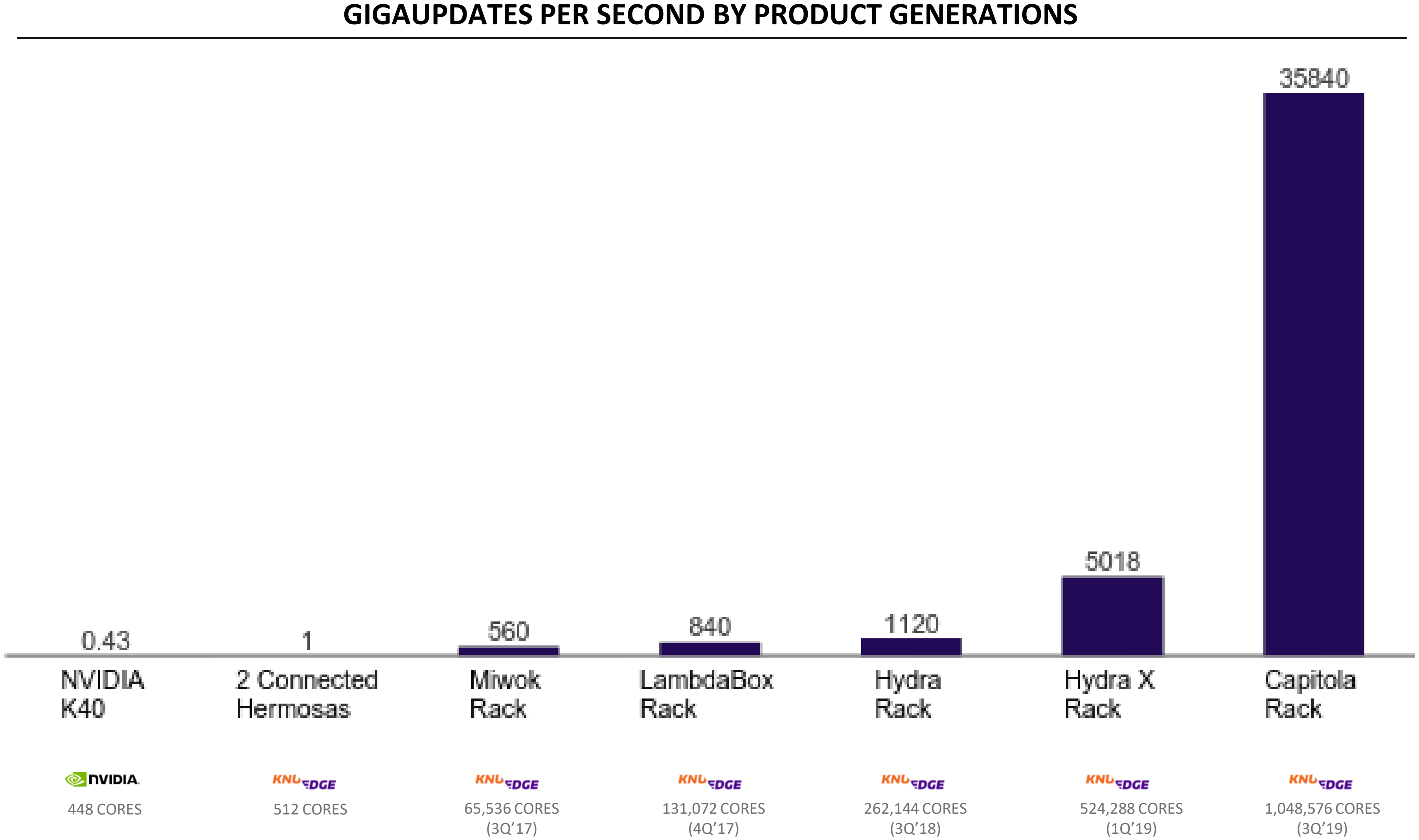
“MOST IMPORTANT BENCHMARK FOR HIGH PERFORMANCE COMPUTING (HPC)”

- GENERATE A RANDOM ADDRESS
- RETRIEVE A VALUE FROM THAT ADDRESS
- XOR THE VALUE WITH A CONSTANT
- WRITE BACK TO THE ADDRESS



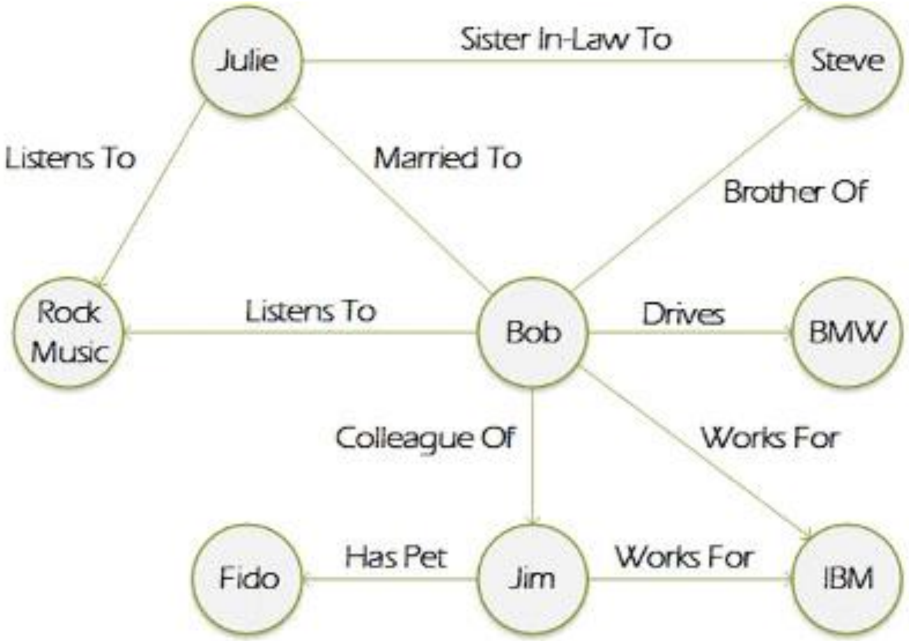


GUPS BENCHMARKING



GRAPH PROCESSING

- Green Graph 500 “Big Data” leader (0.0629 GTEPS/W, 30-scale graph)
- “Small Data” processor that can accommodate 27-scale graphs or larger (0.031 GTEPS/W)



Exceeds small and large leaders by a factor of 500 to over 1000

GRAPH PROCESSING COMPARISON

Add context

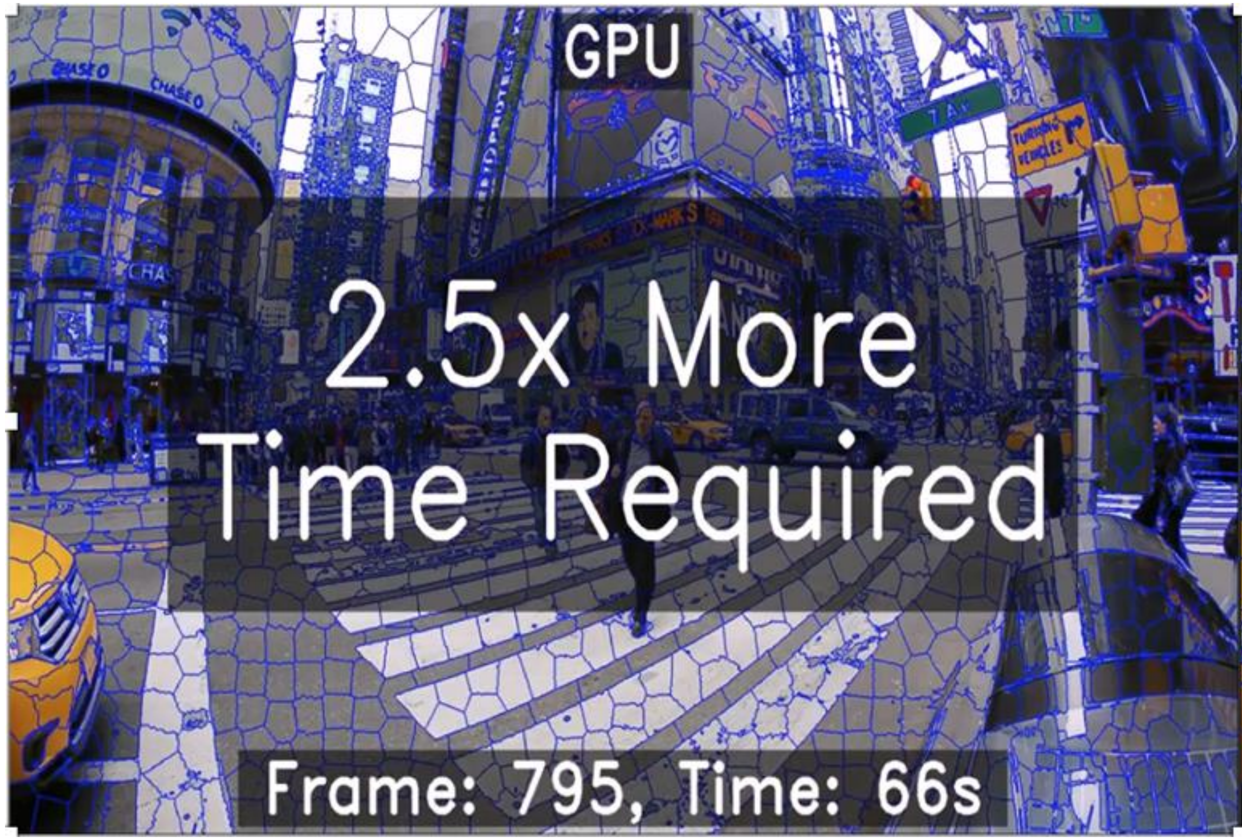
Now	2^{32} Vertices MTEPS/W	2^{32} Vertices GTEPS
Current Record	**52**[1]	**200**[2]
Hermosa	**228**	**44K**

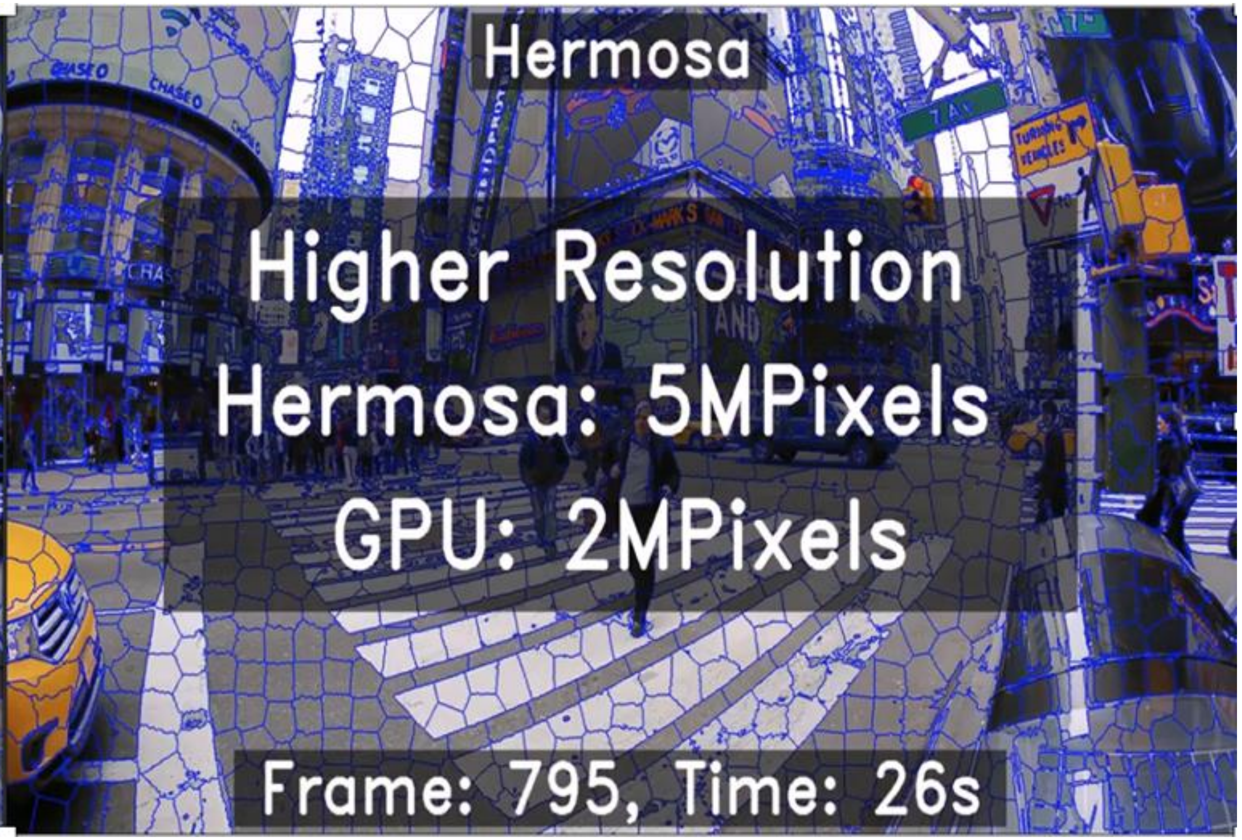
Future	2^{40} Vertices MTEPS/W	2^{40} Vertices GTEPS
Current Record	**4.4**[3]	**39K**[4]
Capitola	**3,400**	**1,400K**

[1] GraphCREST-Huawei, Kyushu University (source: GreenGraph.org, June 2016)
[2] HA-PACS, Center for Computational Sciences, University of Tsukuba (source: Graph500.org, November 2016)
[3] DOE/SC/ANL Mira, Argonne National Laboratory (source: GreenGraph.org, June 2016)
[4] Fujitsu K Computer, RIKEN Advanced Institute for Computational Science (source: Graph500.org, November 2016)
[5] Insufficient address space

IMAGE SEGMENTATION / K-MEANS CLUSTERING

44 more frames per second than a Titan X-based GPU
Process 2MP for Titan X (27,238 cores) vs. 5MP with 4 KnuEdge chips (1,024 cores)



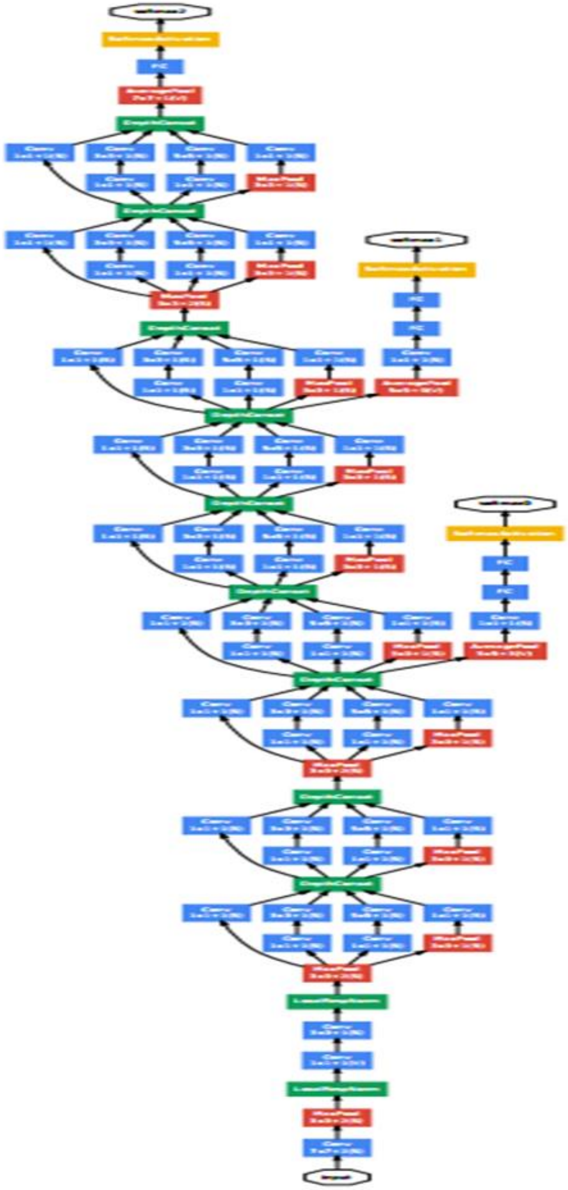


GoogLeNet (CNN)

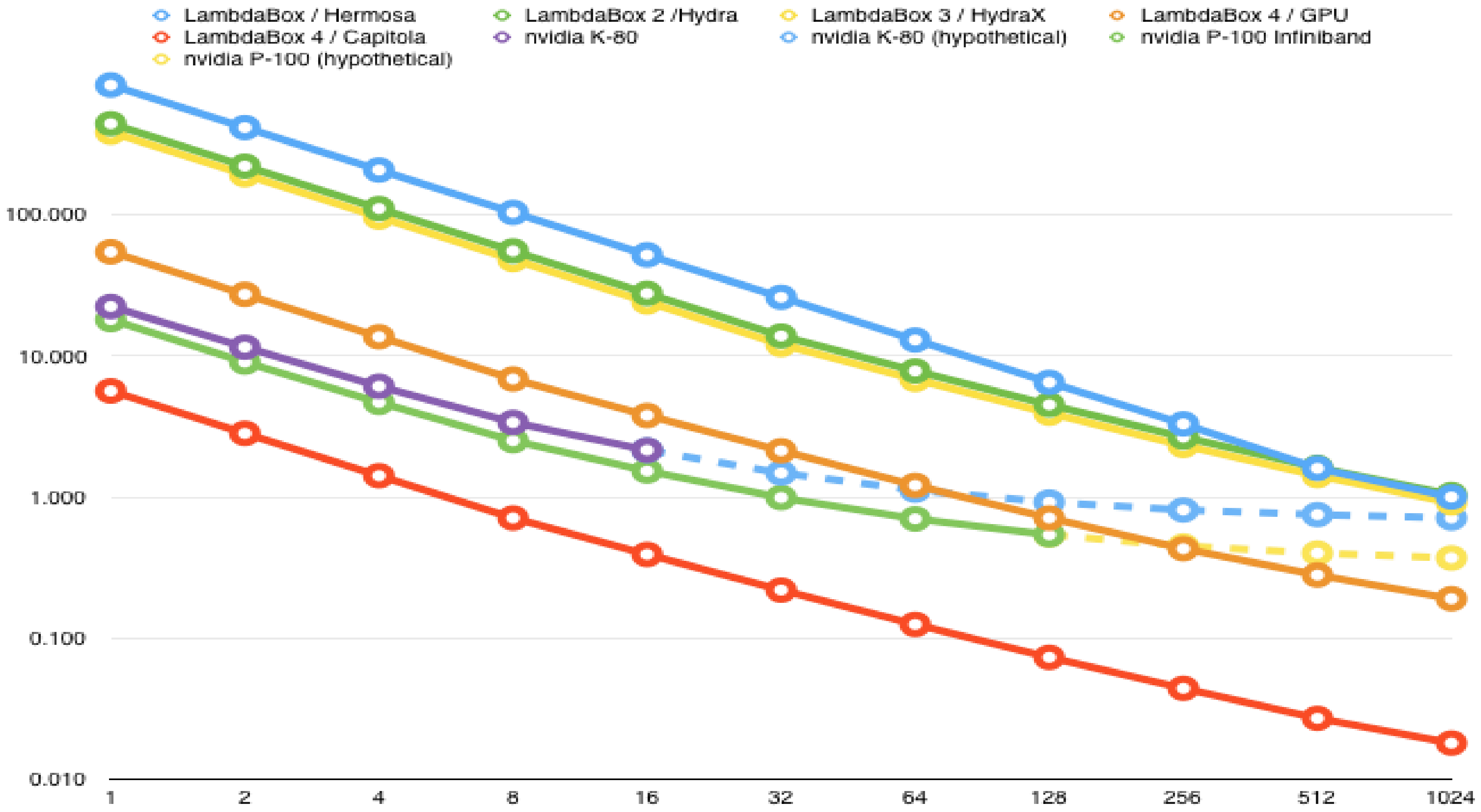
- 1000 categories
- 1.2M images
- 22 layers (27 with pooling)
- 6.8M parameters
- 1.5B operations

(a) Siberian husky

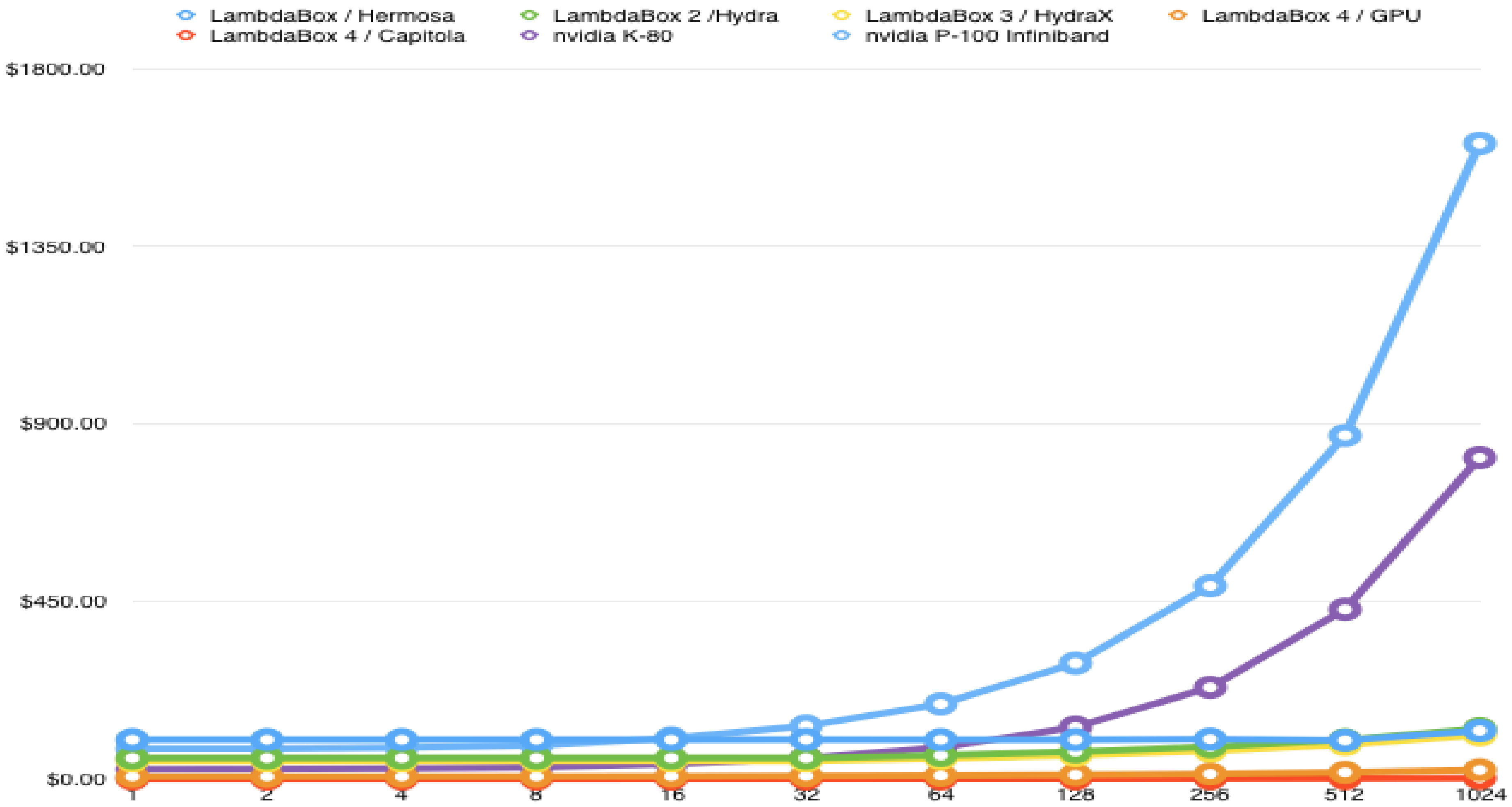
(b) Eskimo dog



TRAINING GOOGLENET - TIME

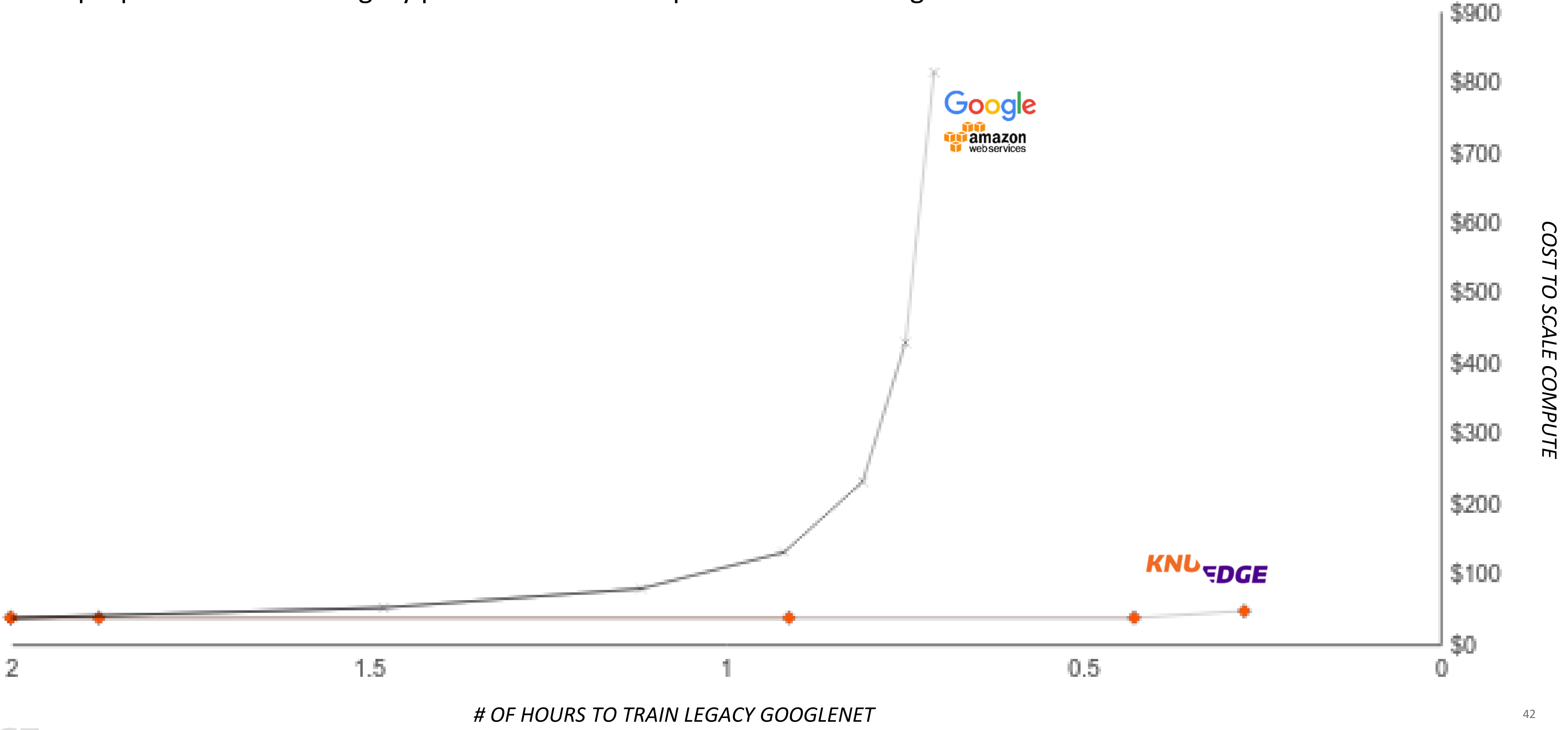


TRAINING GOOGLENET - COST



AT A FLAT COST, KNUEDGE CAN MAKE TRAINING CHALLENGES OBSOLETE

Value proposition even for legacy problems like CNN optimized for existing architectures

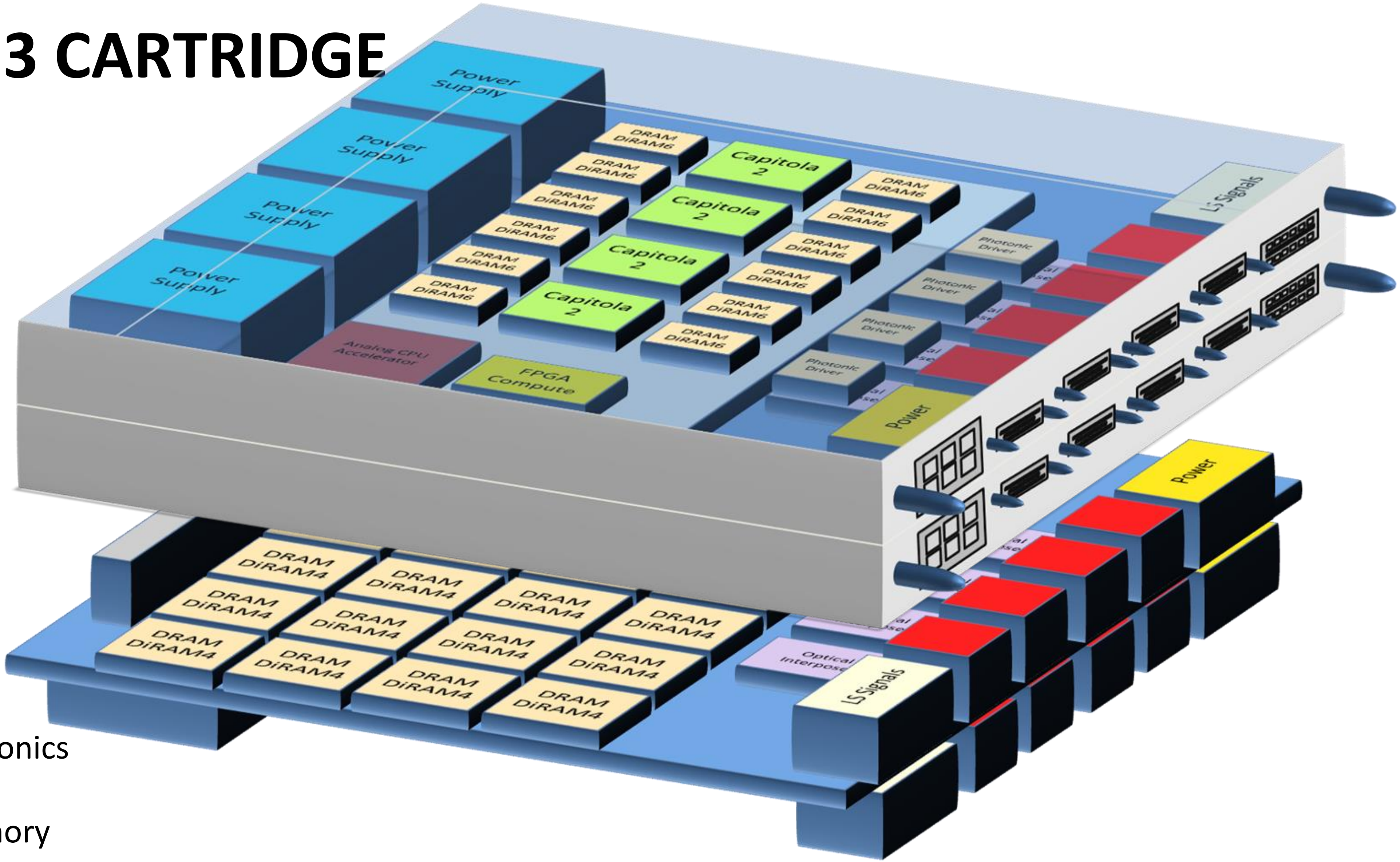


THE MOST POWERFUL SUPERCOMPUTER ACCESSIBLE BY EVERYONE

	TaihuLight (Largest supercomputer today)	**KnuEdge** (By Q4 2018[1])	**KnuEdge** (By Q4 2019[1])	**KnuEdge** (By Q4 2020[1])
CORES	10,649,600	10,600,192	25,165,824	134,217,728
PROCESSOR	Sunway SW26010 1.45GHz	KnuEdge Hermosa 1.0GHz	KnuEdge Capitola 2.0GHz	KnuEdge Capitola 2.0GHz
LINPACK (RMAX) PERFORMANCE	105 PFlop/s	150 PFlop/s	415.2 PFlop/s	2.21 EFlop/s
POWER	15.4 MW	5.1 MW	3.7 MW	19.8 MW
MEMORY	1.31 PB (5.591 PB/s BW)	642 TB (6.625 PB/s BW)	12.8 PB (15.72PB/s BW)	68.1 PB (83.89PN/s BW)
INTERCONNECT	Sunway	LambdaFabric	LambdaFabric	LambdaFabric
OPERATING SYSTEM	Sunway RaiseOS 2.0.5	Linux	Linux	Linux
COST	1.8BYuan (US$273M)	US$100M estimated	US$49.6M estimated	US$264.6M estimated

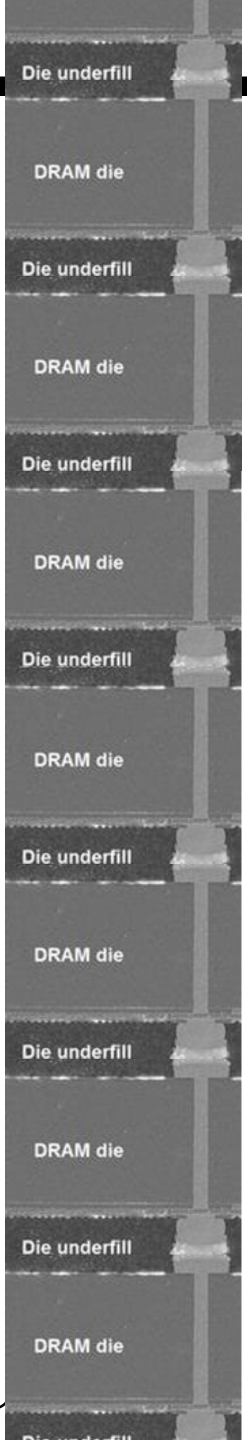
[1] Deployment timing based on demand forecast as of April 16, 2017.

GEN3 CARTRIDGE

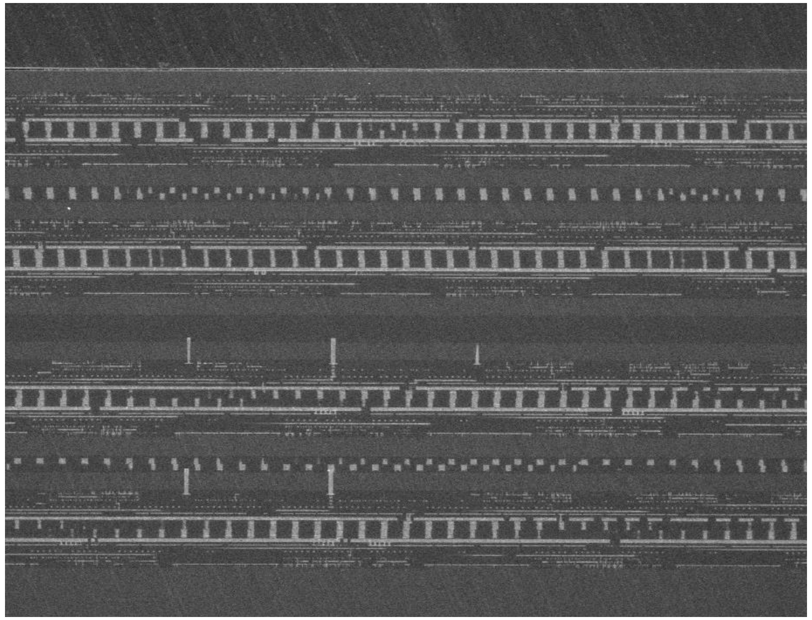


Integrated Photonics
110x100x25mm
>1TB Near Memory
>100TFLOPS

INGREDIENTS

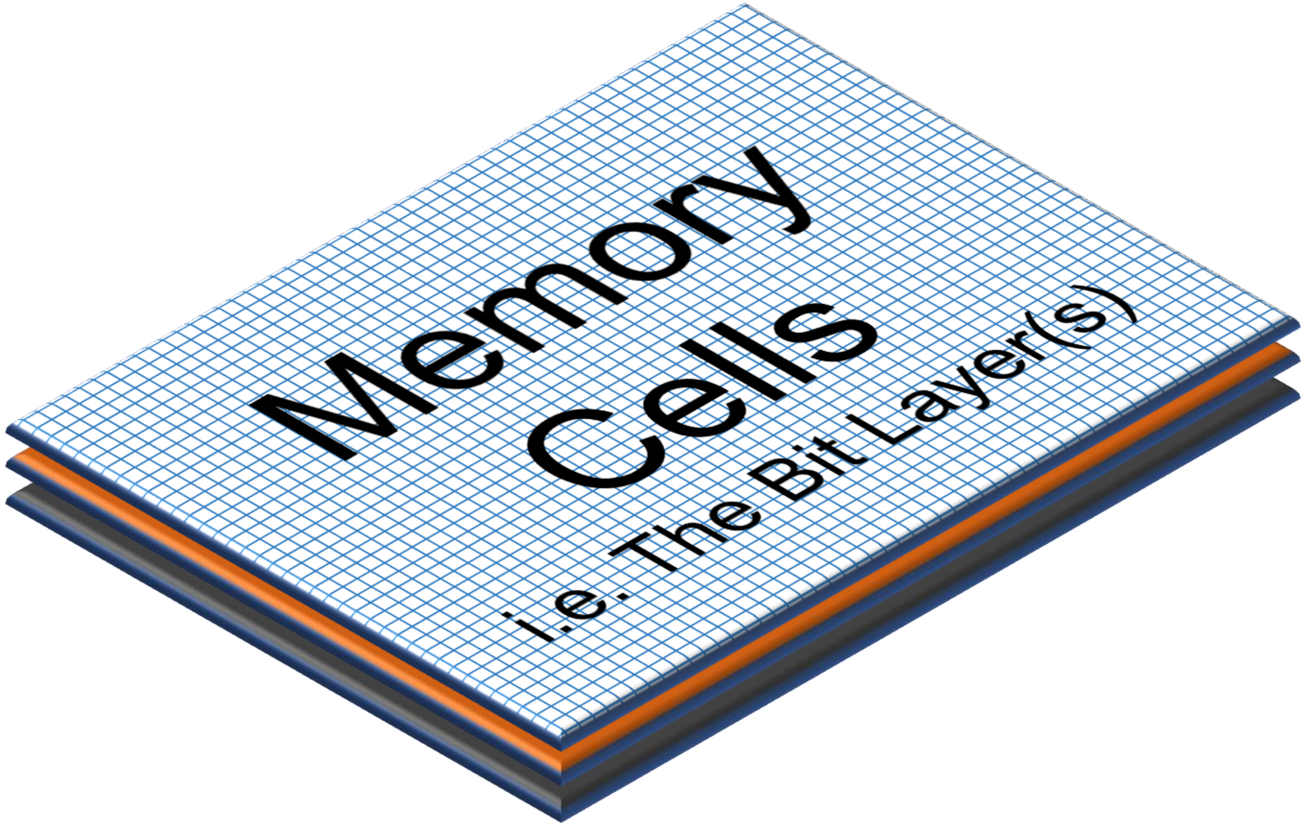


TRUE ARCHITECTURAL 3D

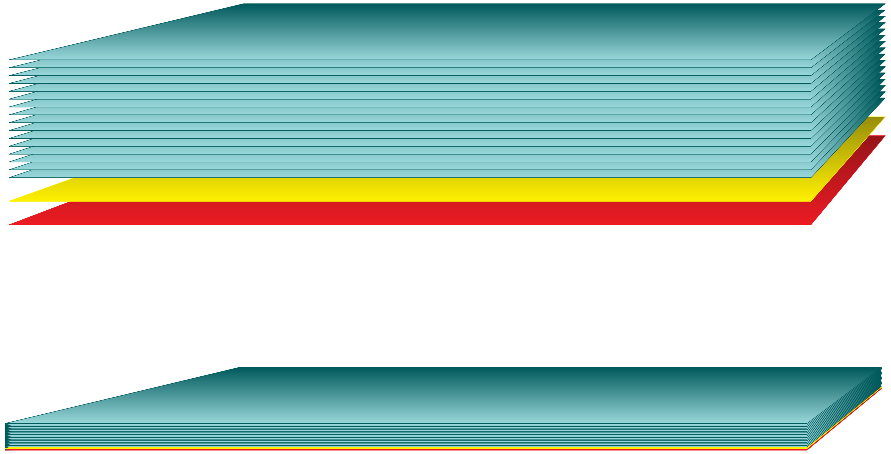


- 8 Layer Stack
- 80 Layers Copper Interconnect
- 2.4 micron Vertical Interconnect Pitch

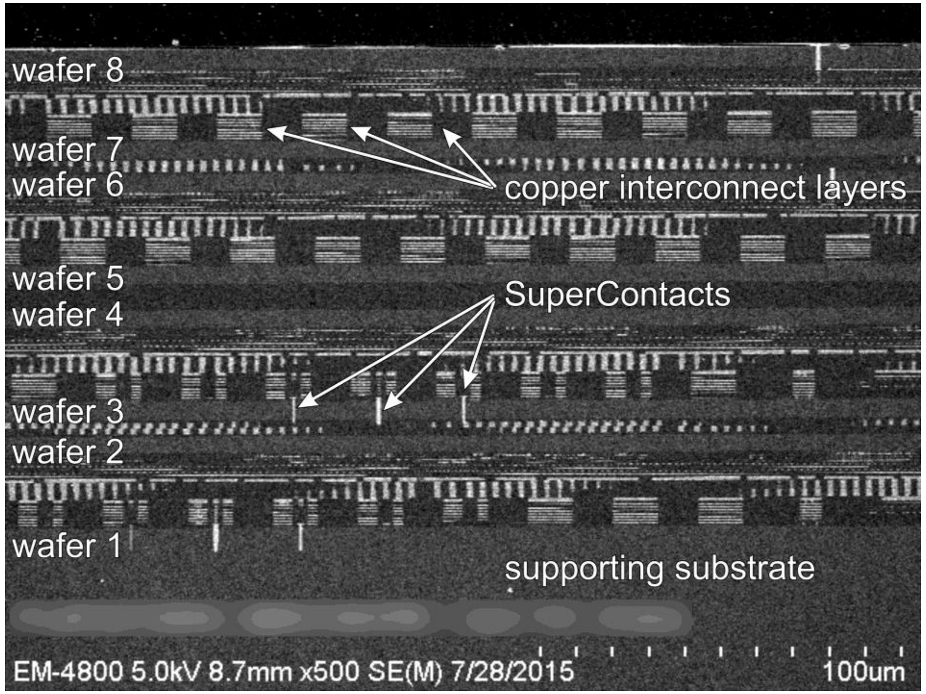
OPTIMIZED – POWER, LATENCY, BANDWIDTH, DENSITY



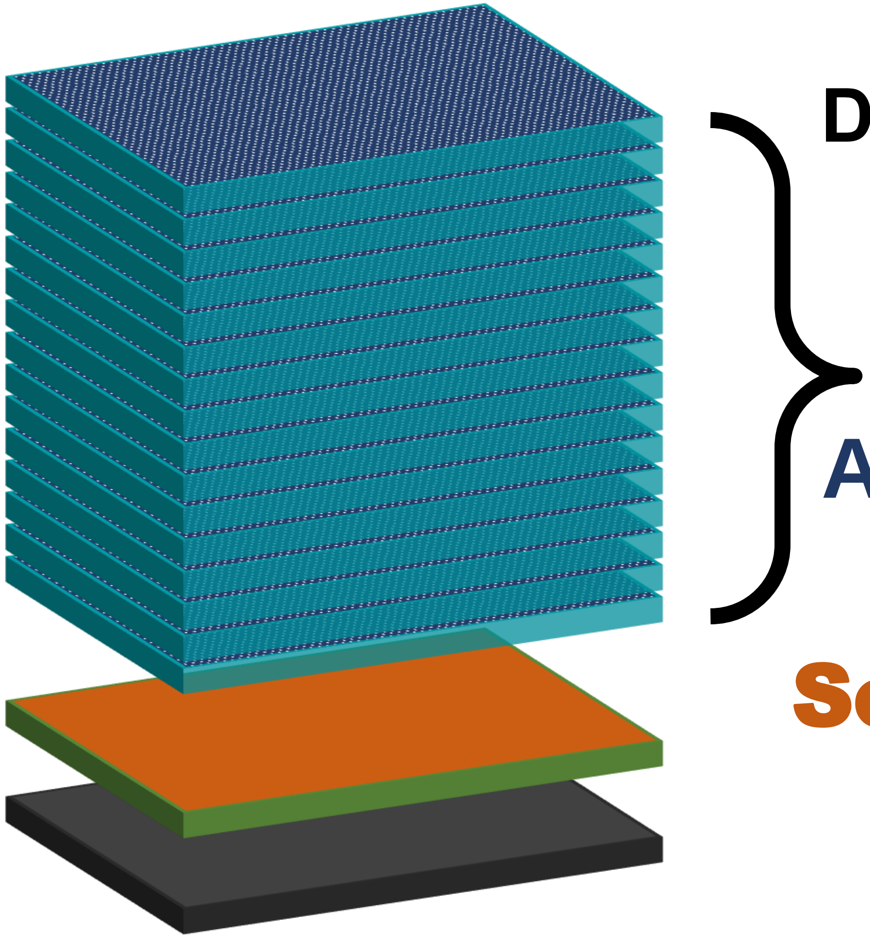
DiRAM4 STACKED FOR PERFORMANCE



- **64 Gb** of Memory in 175 mm^2
- **256** fully independent RAMs
- **16 Banks** per RAM
  - **4096** Banks per device
- **64 bit** Sep I/O Data per RAM
- **7/9ns Access Time** (Closed page to data)
- **15ns tRC** (Page Open to Page Open in a Bank)
- 16.4 Terabit/s Data Bandwidth
- **>500** Billion Transactions Per Second (Minimum)



DIS-INTEGRATED 3D RAM ARCHITECTURE



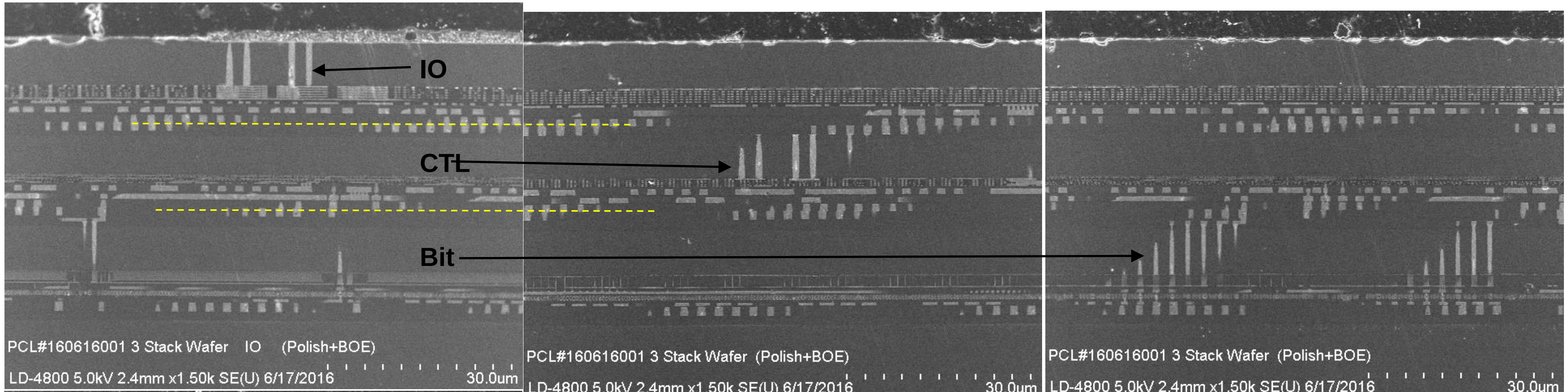
DiRAM™ Architecture

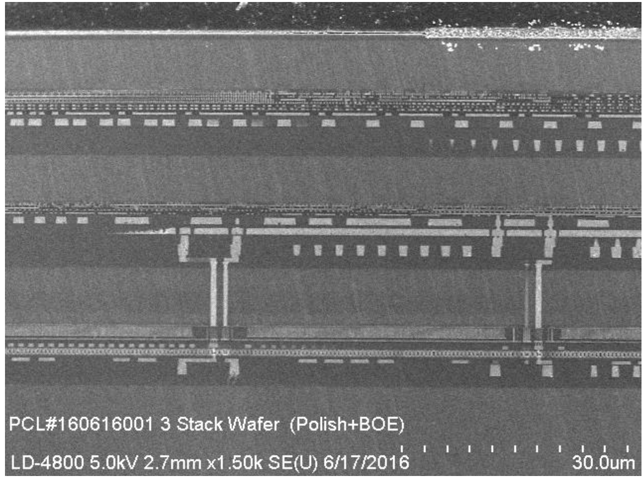
Memory Cells
and
Access Transistors

Sense Amps, etc.

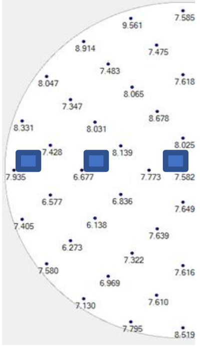
I/O Layer

DiRAM CLOSE-UP

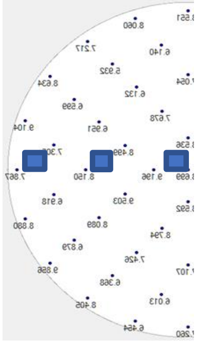

IO
CTL
Bit
PCL#160616001 3 Stack Wafer IO (Polish+BOE)
LD-4800 5.0kV 2.4mm x1.50k SE(U) 6/17/2016 30.0um
PCL#160616001 3 Stack Wafer (Polish+BOE)
LD-4800 5.0kV 2.4mm x1.50k SE(U) 6/17/2016 30.0um
PCL#160616001 3 Stack Wafer (Polish+BOE)
LD-4800 5.0kV 2.4mm x1.50k SE(U) 6/17/2016 30.0um


PCL#160616001 3 Stack Wafer (Polish+BOE)
LD-4800 5.0kV 2.7mm x1.50k SE(U) 6/17/2016 30.0um

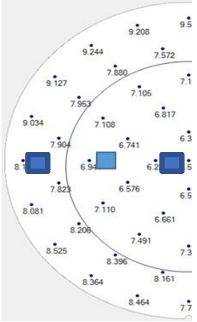
F/A point: 3 points : wafer left-edge, middle and center



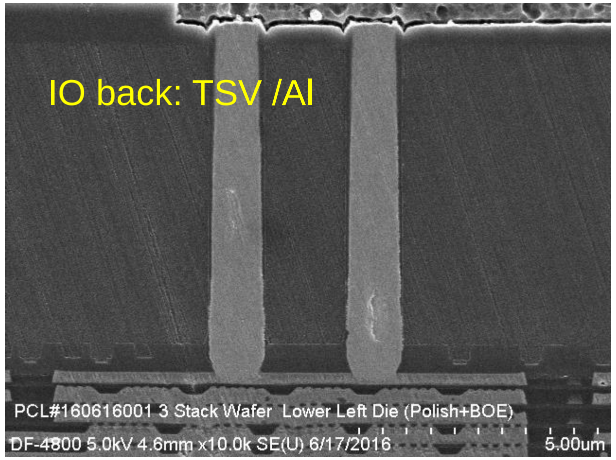
IO Si thickness

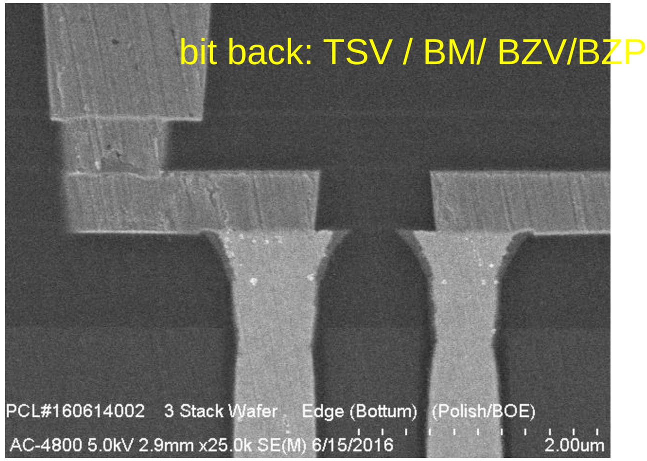


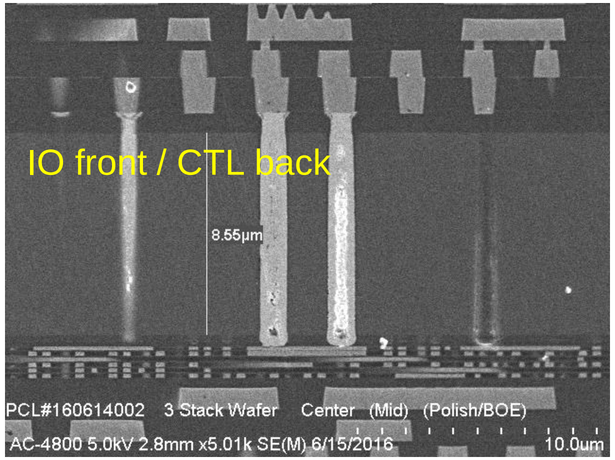
CTL Si thickness

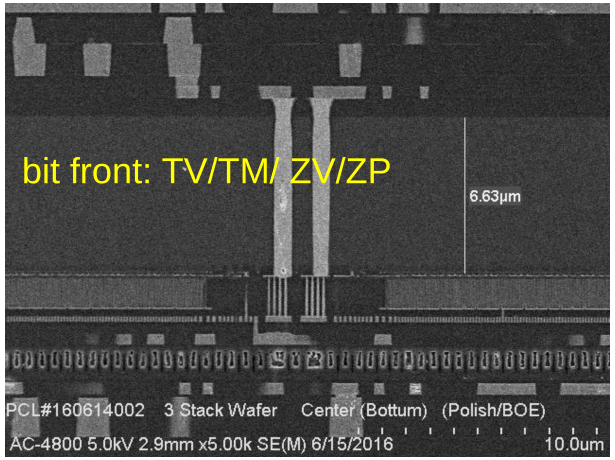


Bit Si thickness


IO back: TSV /Al
PCL#160616001 3 Stack Wafer Lower Left Die (Polish+BOE)
DF-4800 5.0kV 4.6mm x10.0k SE(U) 6/17/2016
5.00um


bit back: TSV / BM/ BZV/BZP
PCL#160614002 3 Stack Wafer Edge (Bottum) (Polish/BOE)
AC-4800 5.0kV 2.9mm x25.0k SE(M) 6/15/2016
2.00um


IO front / CTL back
8.55µm
PCL#160614002 3 Stack Wafer Center (Mid) (Polish/BOE)
AC-4800 5.0kV 2.8mm x5.01k SE(M) 6/15/2016
10.0um


bit front: TV/TM/ ZV/ZP
6.63µm
PCL#160614002 3 Stack Wafer Center (Bottum) (Polish/BOE)
AC-4800 5.0kV 2.9mm x5.00k SE(M) 6/15/2016
10.0um

SUPER-DENSE INTERCONNECT ALLOWS…

Bi-STAR®

Built-**i**n **S**elf **T**est **A**nd **R**epair

- Controlled by **embedded ARM processor**
- Super-fine grained test and repair
- Massively parallel test
- At-speed wafer probe testing
- Post-assembly repair
- Continuous, in-the-system hard and soft error repair

Bi-STAR® DOES MORE, WORKS BETTER

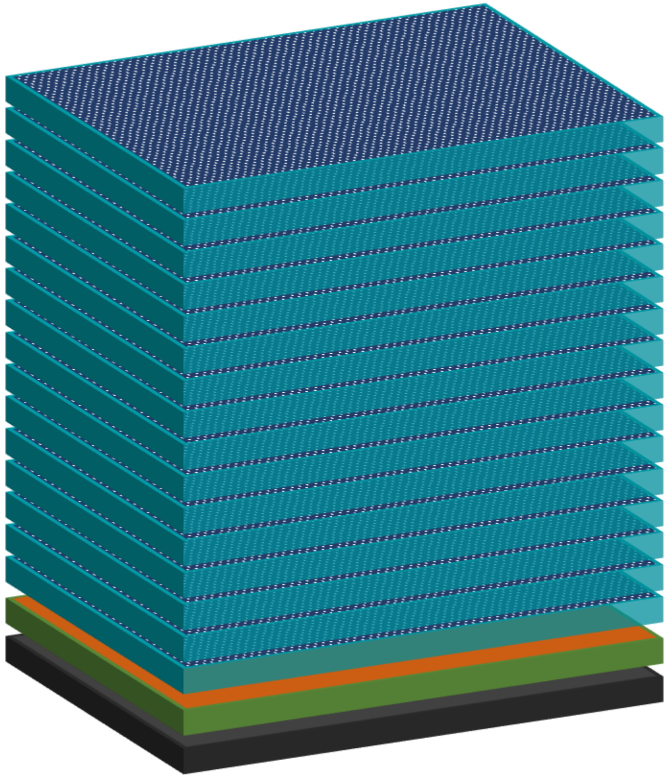
Bi-STAR Repairs

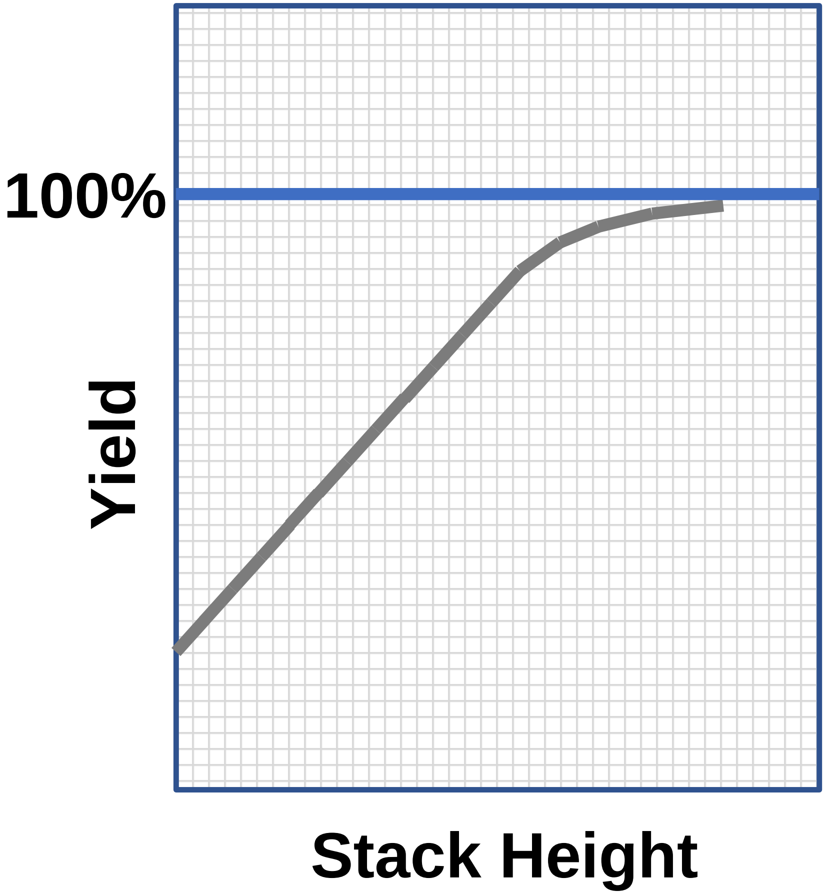
- Bad memory cells
- Bad line drivers
- Bad sense amps
- Shorted word lines
- Shorted bit-lines
- Leaky bits
- Bad secondary bus drivers

Bi-STAR Tests

- Tests > 300,000 nodes per clock cycle
- Tests > 1,000x faster than external memory tester
- Via SPI, I2C port, or JTAG. Works with Host to allow continuous scrub / repair

Bi-STAR REPAIR IMPROVES YIELD



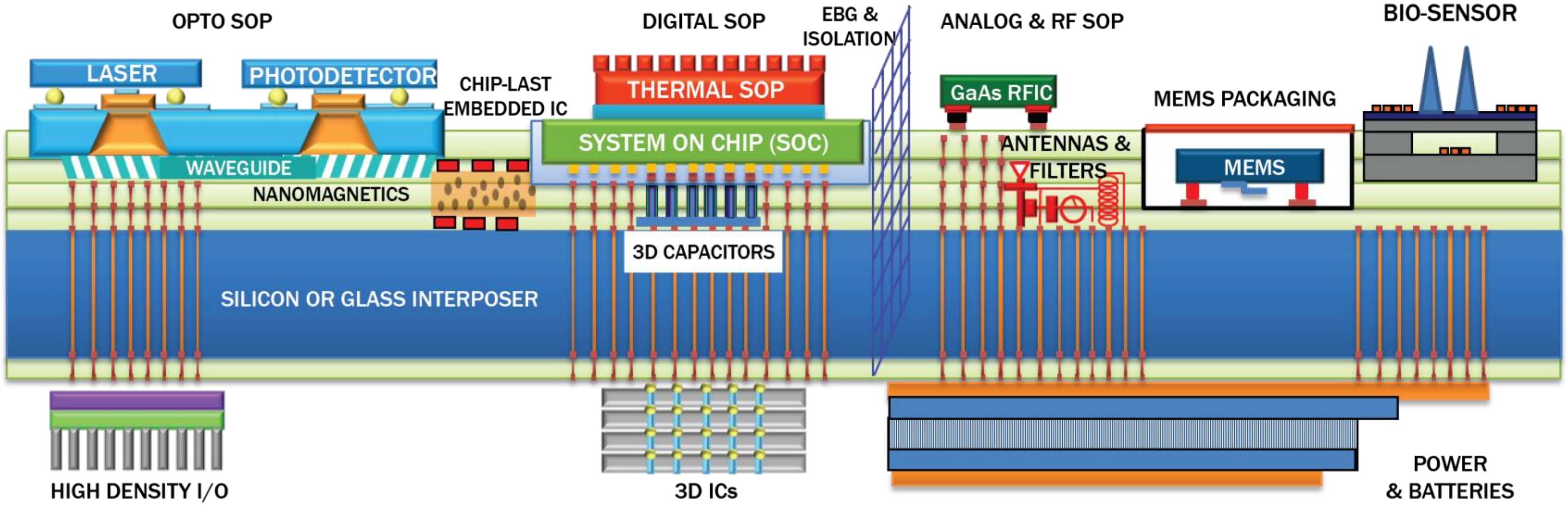


VERY ADVANCED PACKAGING

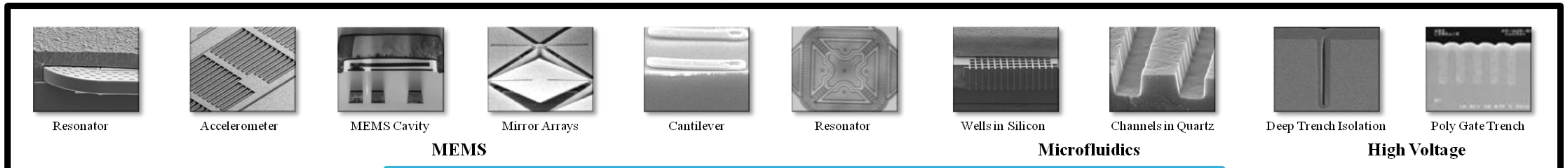
MORE THAN MOORE

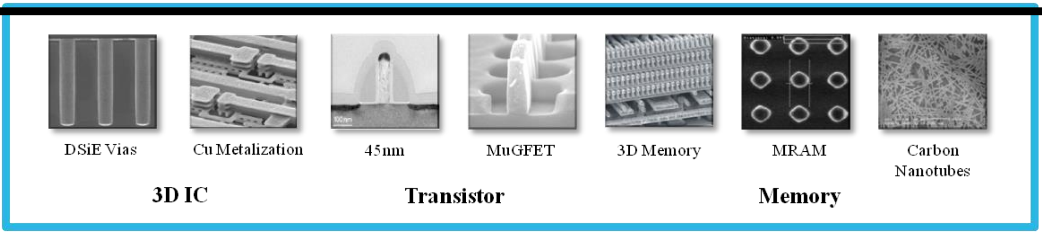
GEORGIA TECH PRC

1. THIN 2D AND 3D ACTIVES
2. THIN FILM PASSIVES
3. SYSTEM INTERCONNECTIONS
4. THERMAL INTERFACES AND STRUCTURES

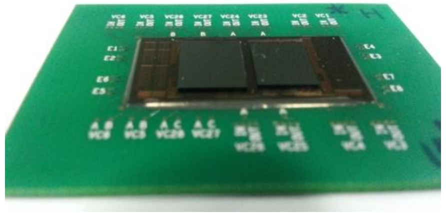


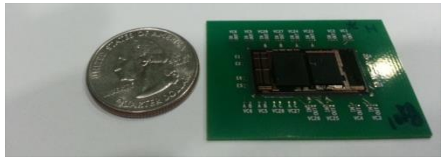
5. MULTI-FUNCTION MATERIALS
6. MIXED SIGNAL DESIGN AND TEST
7. MECHANICAL DESIGN AND RELIABILITY
8. POWER SOURCES

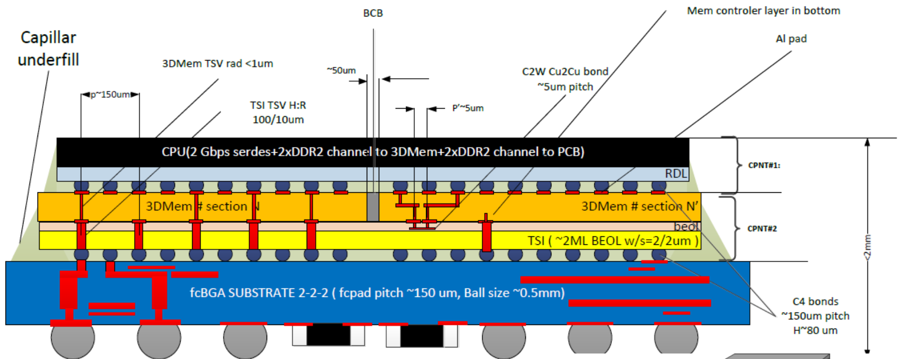




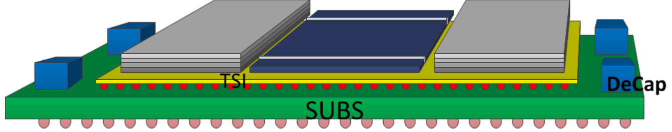
2.5/3D CIRCUITS

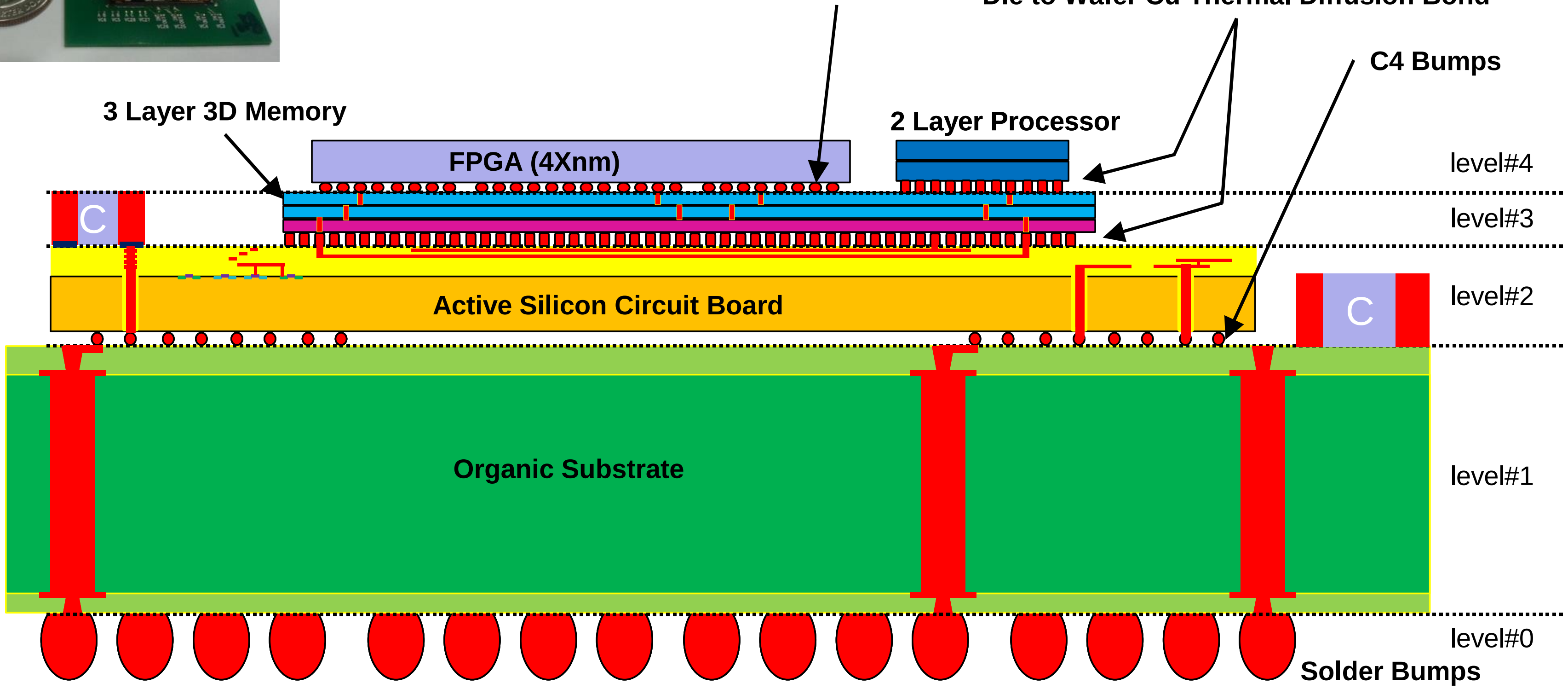






BCB
Mem controler layer in bottom
Capillar underfill
Al pad
3DMem TSV rad <1um
~50um
C2W Cu2Cu bond ~5um pitch
p~150um
TSI TSV H:R 100/10um
P'~5um
CPU(2 Gbps serdes+2xDDR2 channel to 3DMem+2xDDR2 channel to PCB)
RDL
CPNT#1:
3DMem # section N
3DMem # section N'
beol
CPNT#2
TSI (~2ML BEOL w/s=2/2um)
<2mm
fcBGA SUBSTRATE 2-2-2 (fcpad pitch ~150 um, Ball size ~0.5mm)
C4 bonds ~150um pitch H~80 um





µBumps

Die to Wafer Cu Thermal Diffusion Bond

C4 Bumps

3 Layer 3D Memory

2 Layer Processor

FPGA (4Xnm)

level#4

level#3

C

Active Silicon Circuit Board

level#2

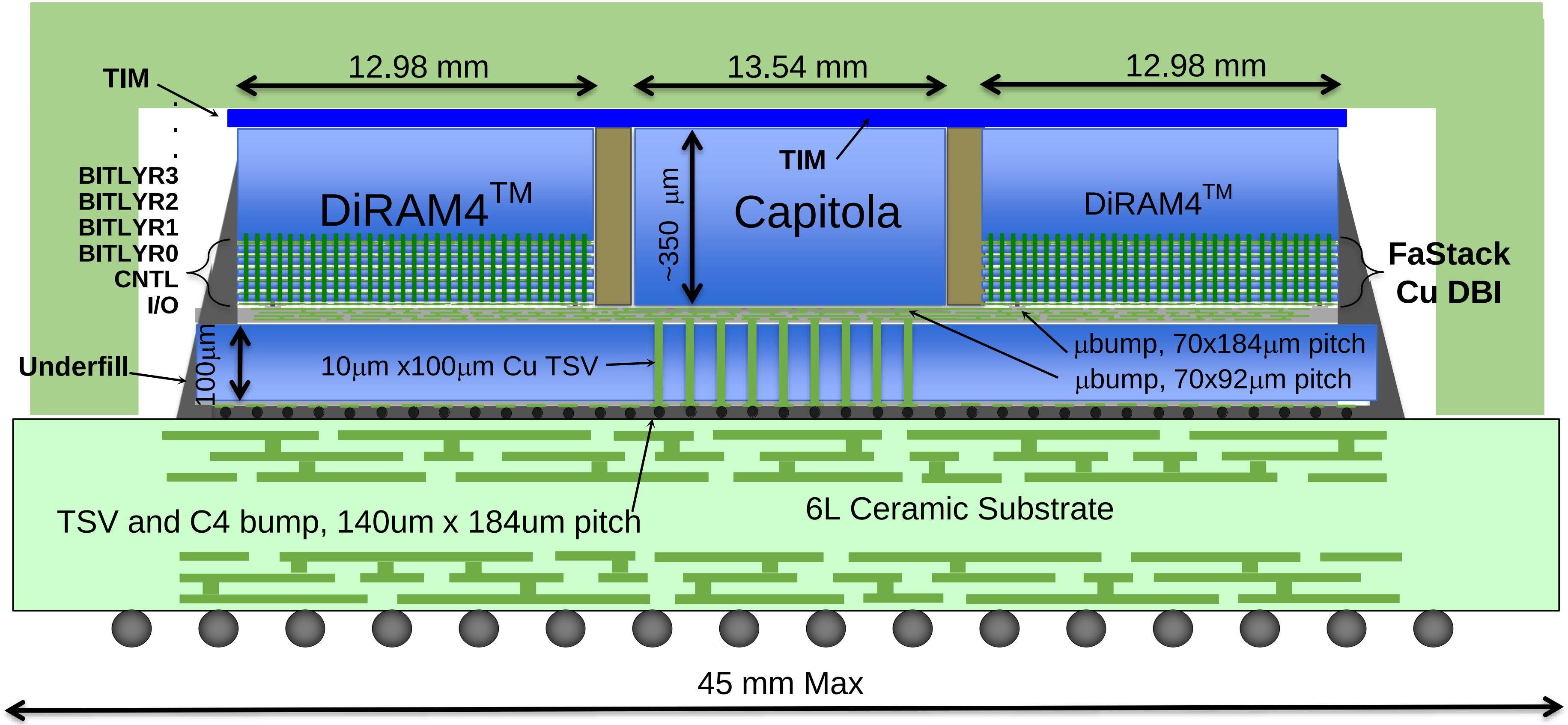
Organic Substrate

level#1

level#0

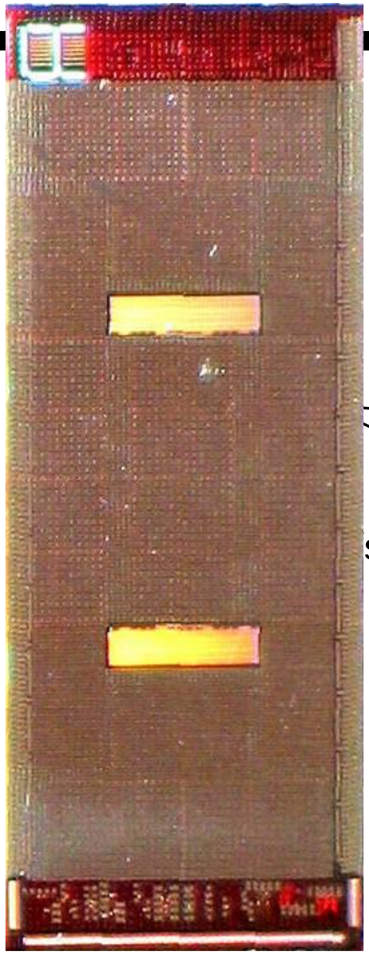
Solder Bumps

MEMORY BLOCK PACKAGE



LOGIC ON MEMORY

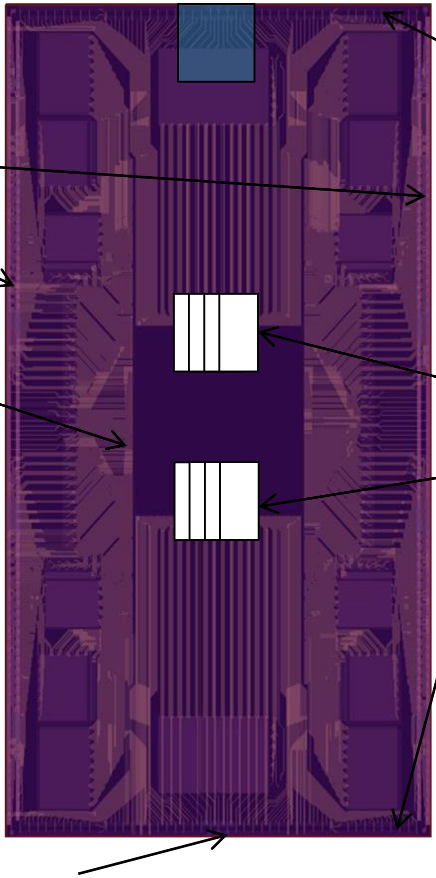
Memory also acts as interposer



Logic

side

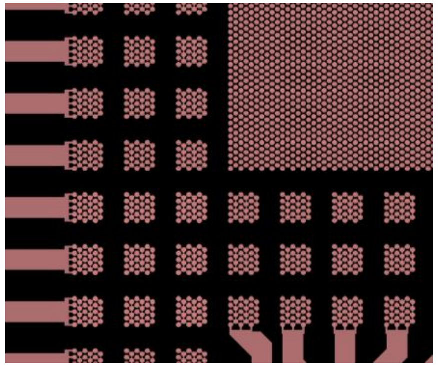
Memory

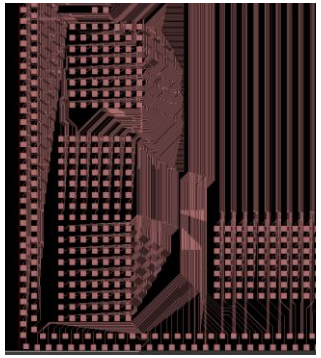


92 pads
(528 total pads at edge,
stagger 250um pad, 125um
pitch
~1500 available pads)

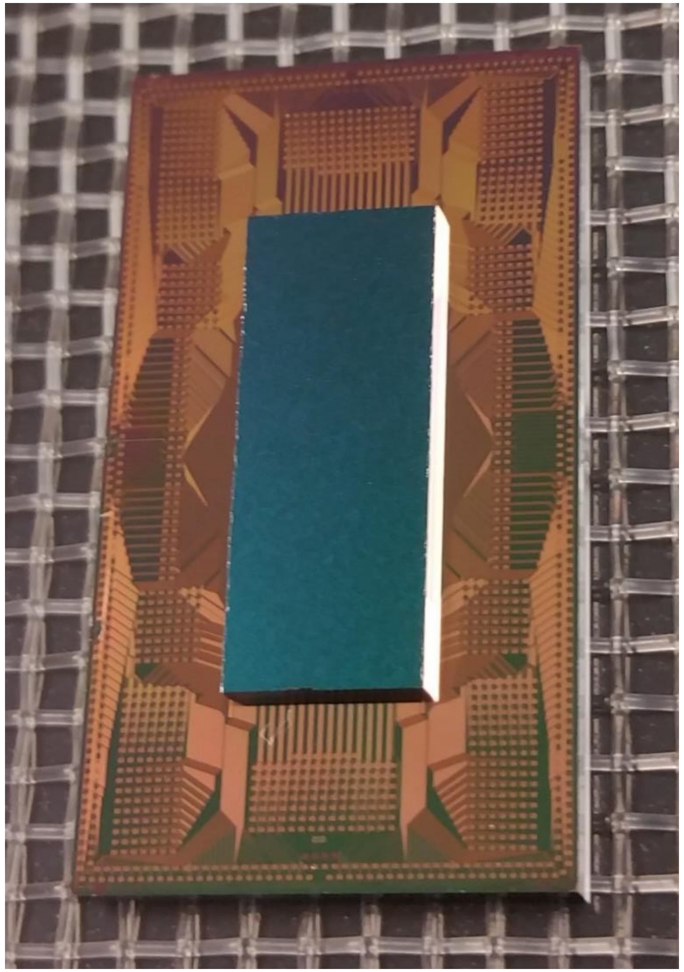
8 DRAM ports
16x21 pad array

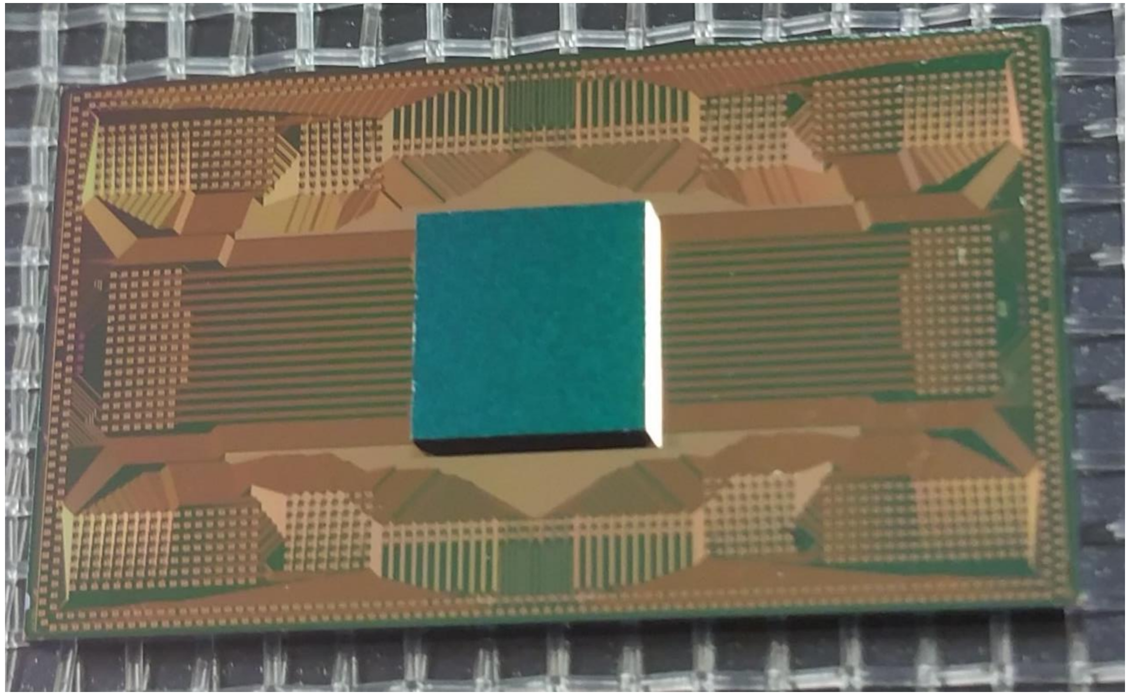
>10µf bypass caps
SS ~4,000pf





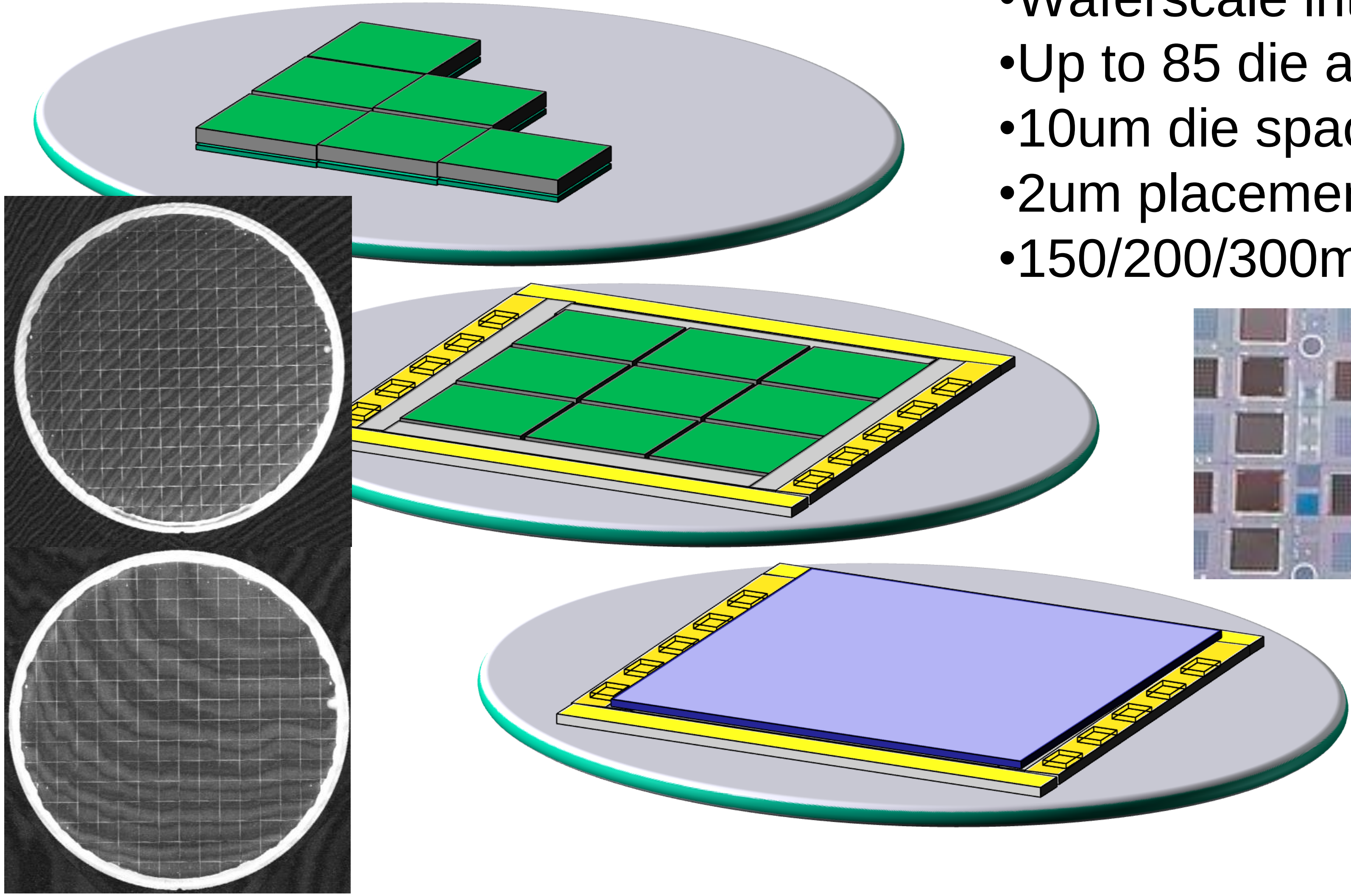
2.5D SYSTEMS IN PRACTICE

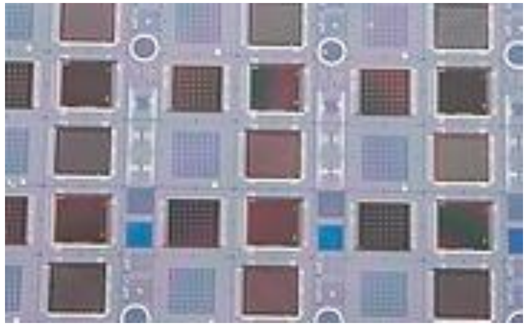




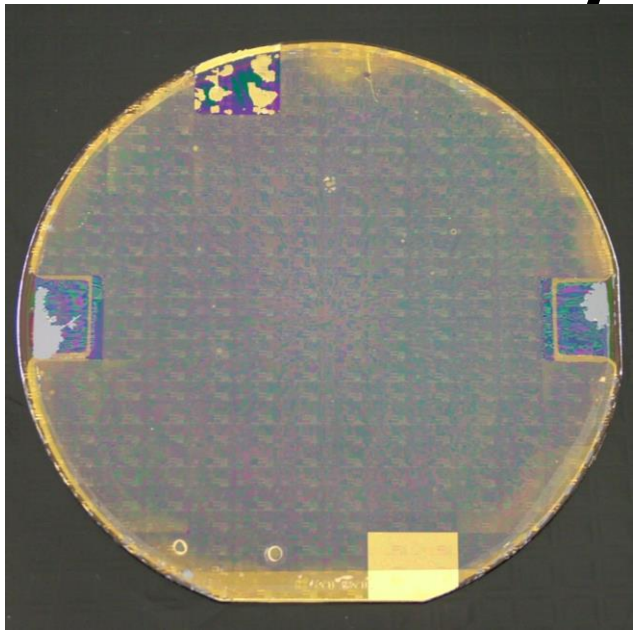
DIE-ON-WAFER AND WAFER-SCALE FPAS

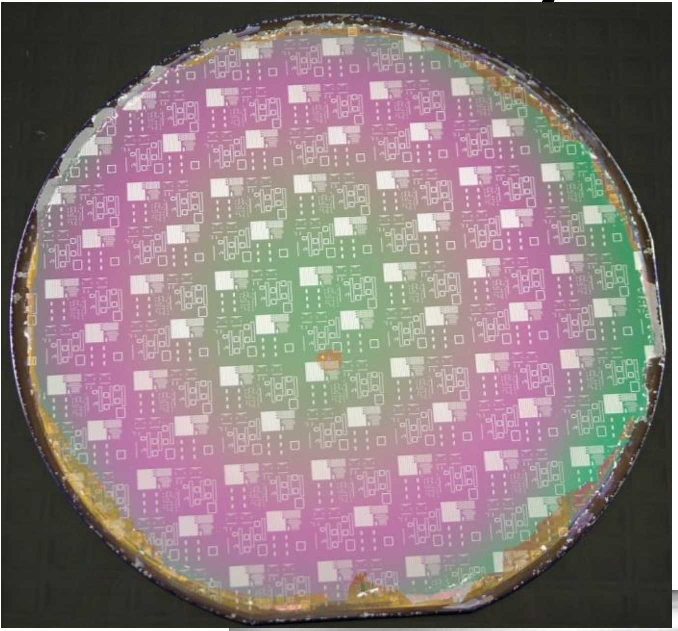
- Waferscale integration
- Up to 85 die assembly
- 10um die space
- 2um placement
- 150/200/300mm



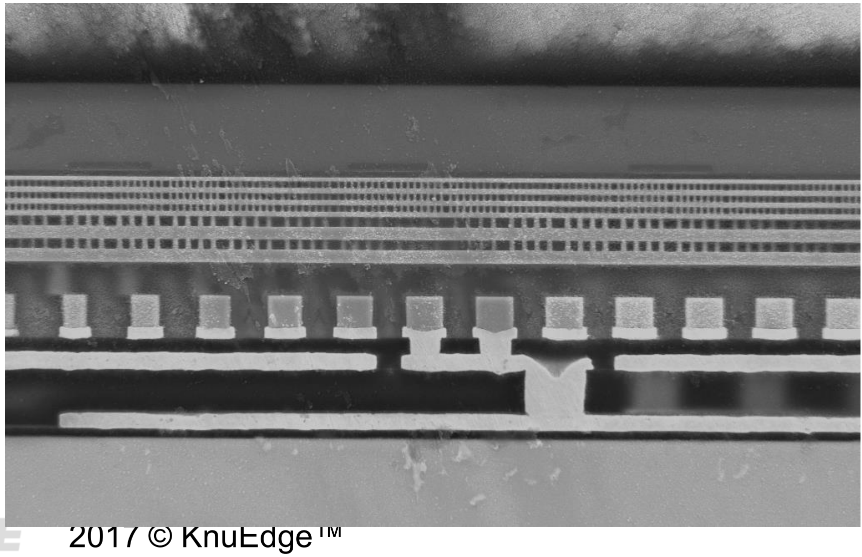


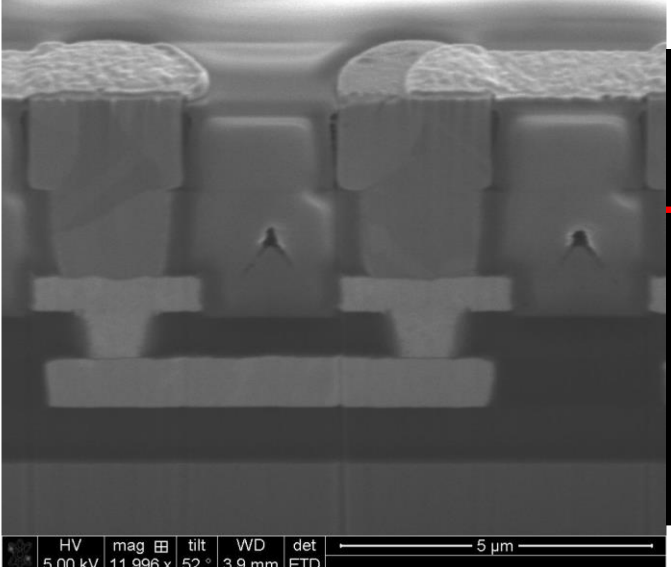
MIXED CMOS-3/5 100MM INP/CMOS

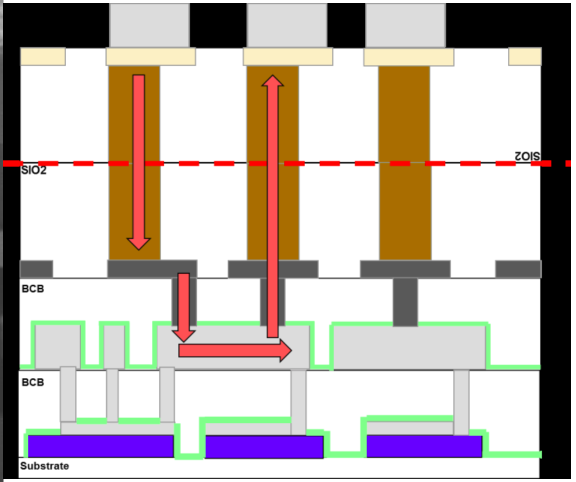




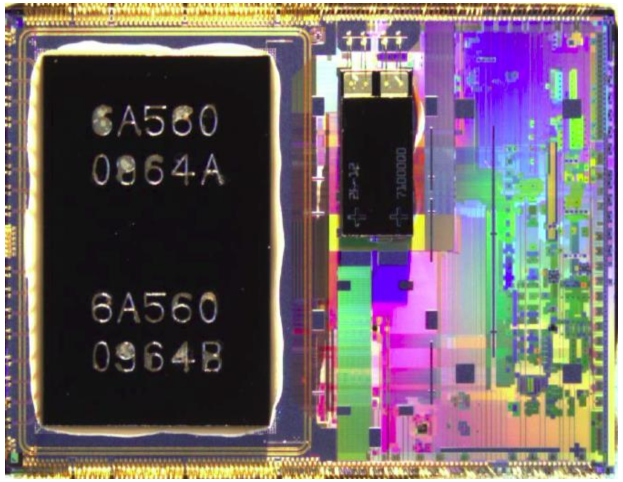
- GaN
- 3D CMOS/InP/GaN
- GaAs
- Graphene

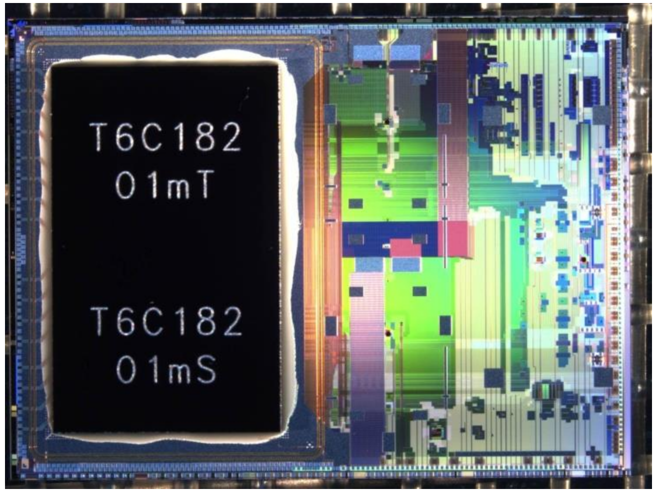






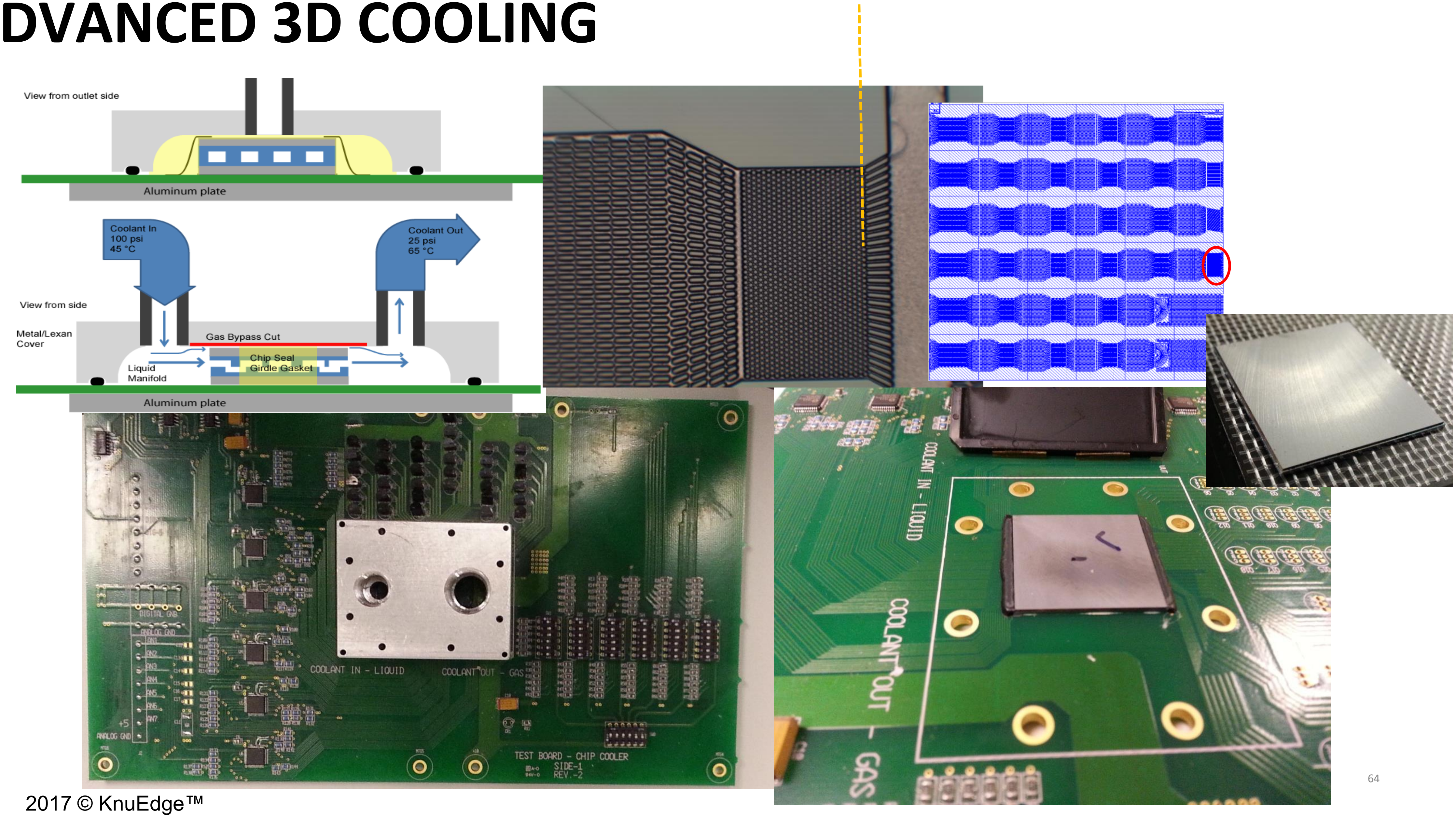
INTERPOSER PHOTONIC DATA PUMP





- Short – Medium Haul
- Module-to-Module
- 8 Core Fiber
- 25Gb → 112Gb
- >1.6Tb/s payload

ADVANCED 3D COOLING


View from outlet side
Aluminum plate
Coolant In
100 psi
45 °C
Coolant Out
25 psi
65 °C
View from side
Metal/Lexan Cover
Gas Bypass Cut
Chip Seal
Girdle Gasket
Liquid Manifold
Aluminum plate
COOLANT IN - LIQUID
COOLANT OUT - GAS
TEST BOARD - CHIP COOLER
SIDE-1
REV.-2
COOLANT IN - LIQUID
COOLANT OUT - GAS