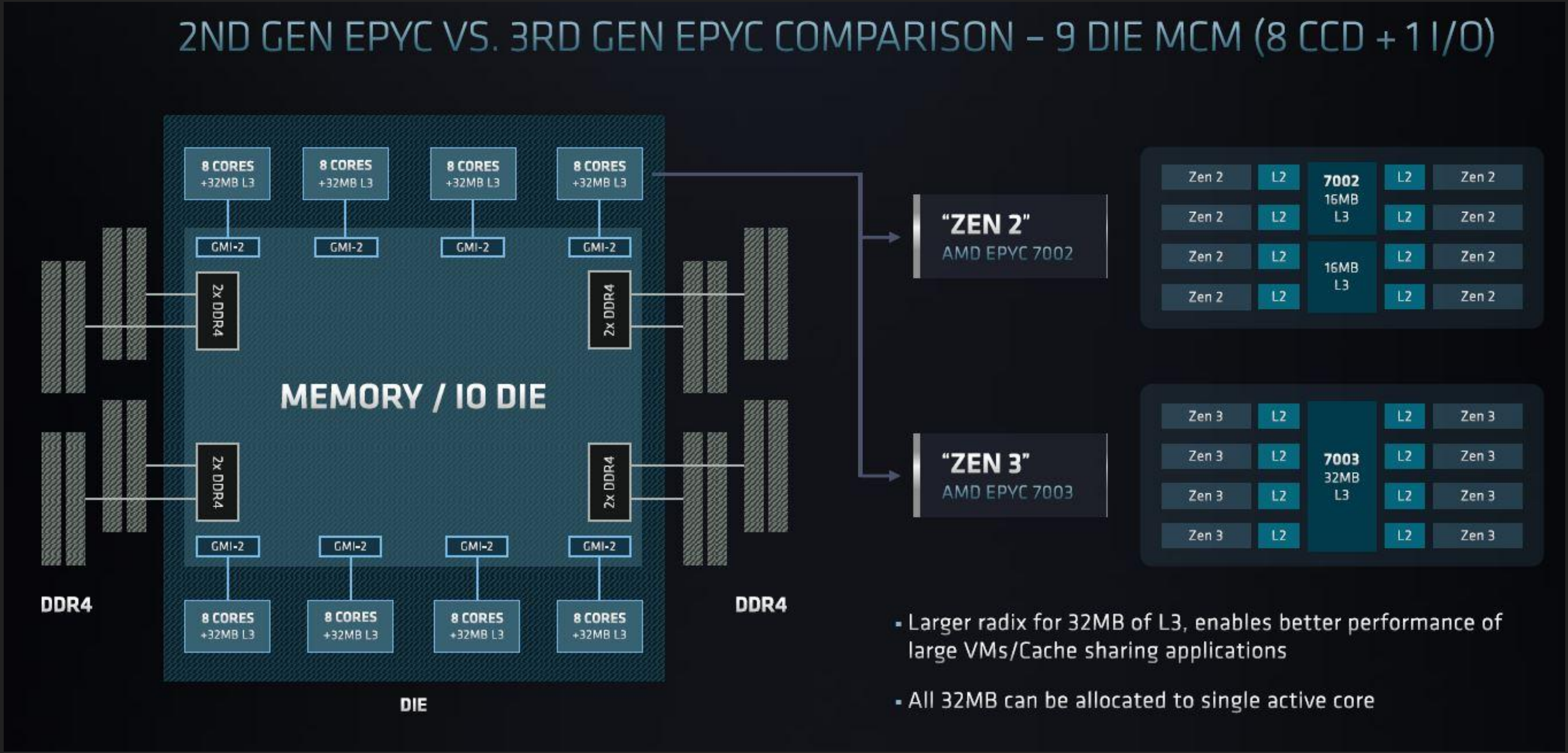
Boston Linux & Unix Users
Boston HPC & GPU

Eliot Eshelman
April 2021

AMD EPYC "Milan"

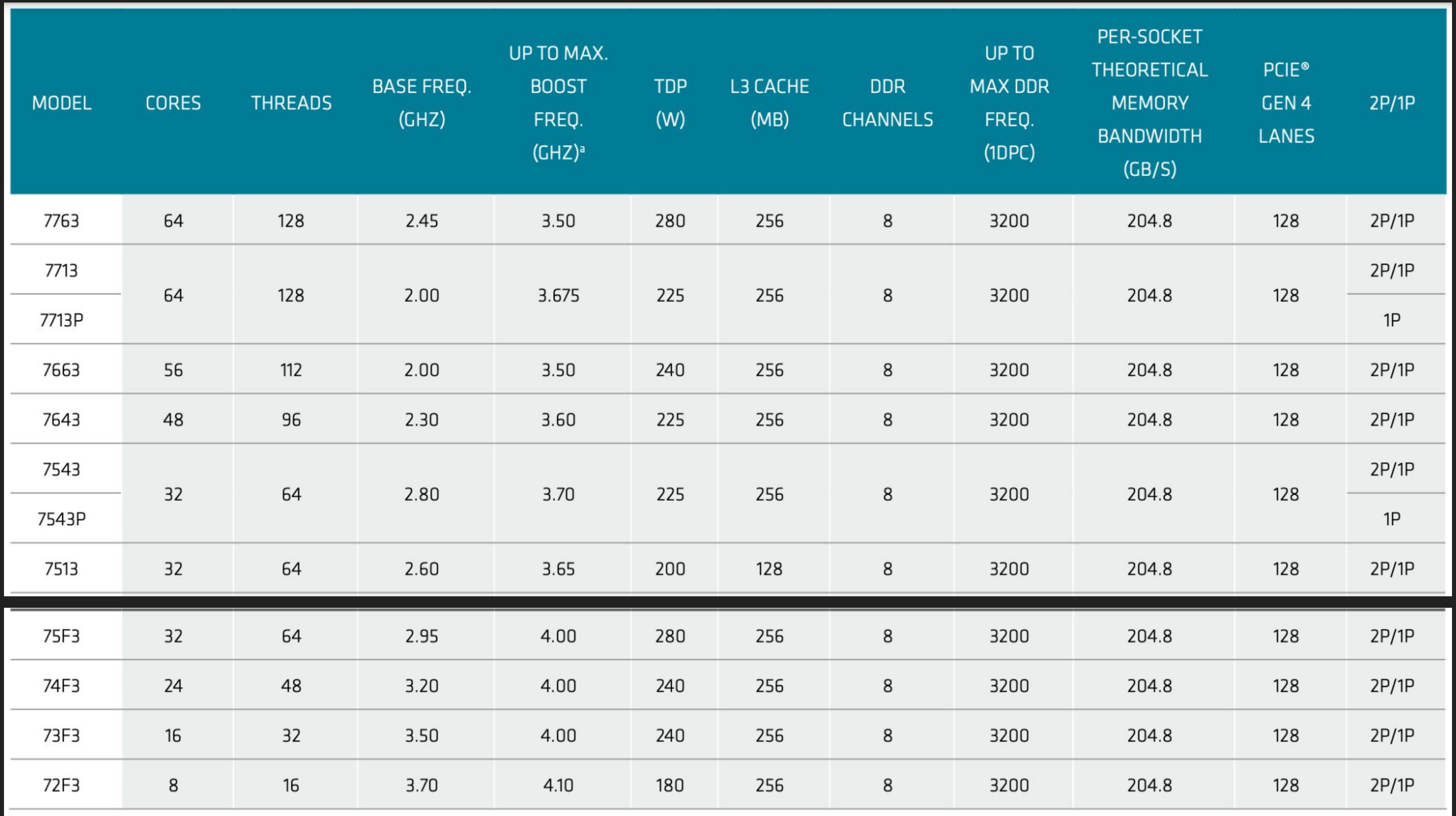
AMD EPYC SOC Architecture (7002- vs 7003-series)



2ND GEN EPYC VS. 3RD GEN EPYC COMPARISON – 9 DIE MCM (8 CCD + 1 I/O)

- Larger radix for 32MB of L3, enables better performance of large VMs/Cache sharing applications
- All 32MB can be allocated to single active core

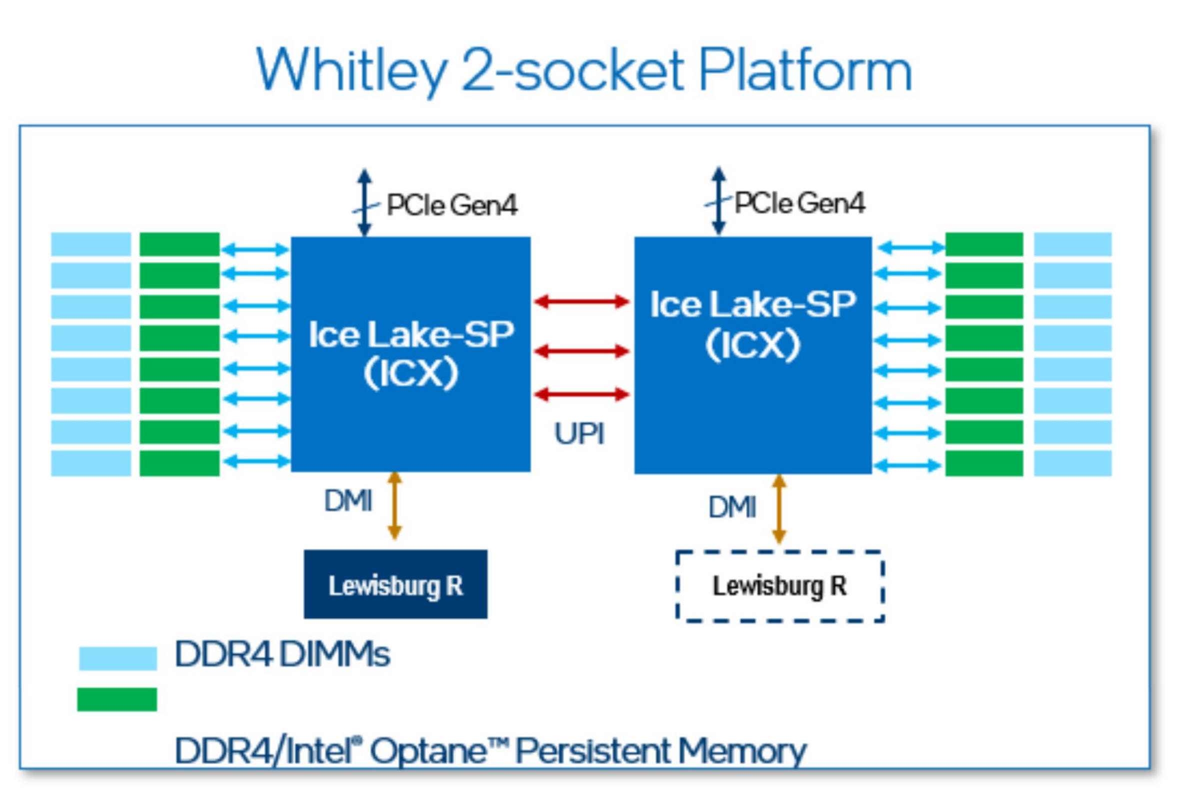
AMD EPYC 7003-series "Milan" CPUs



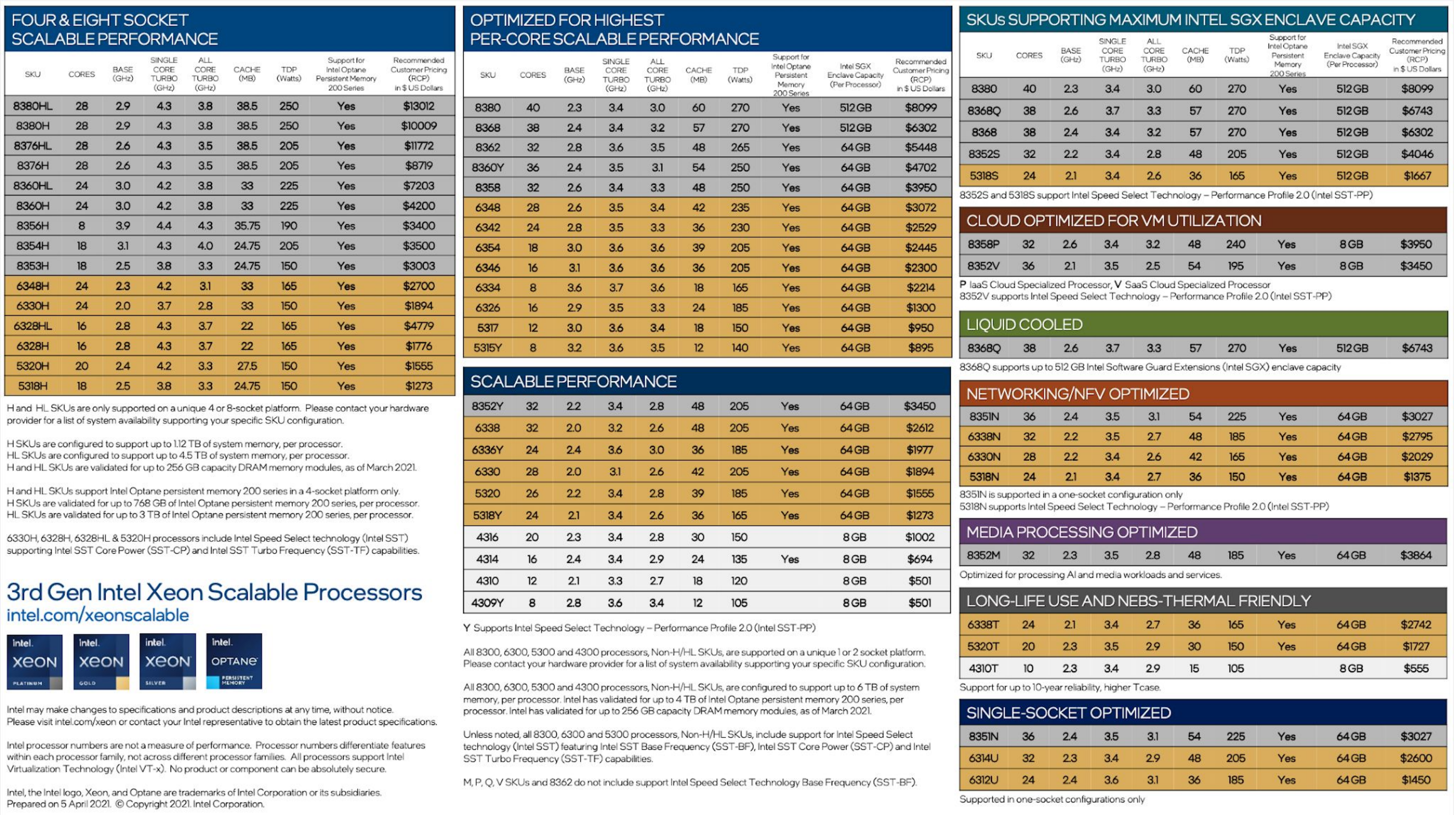
MODEL	CORES	THREADS	BASE FREQ. (GHZ)	UP TO MAX. BOOST FREQ. (GHZ)[a]	TDP (W)	L3 CACHE (MB)	DDR CHANNELS	UP TO MAX DDR FREQ. (1DPC)	PER-SOCKET THEORETICAL MEMORY BANDWIDTH (GB/S)	PCIE® GEN 4 LANES	2P/1P
7763	64	128	2.45	3.50	280	256	8	3200	204.8	128	2P/1P
7713	64	128	2.00	3.675	225	256	8	3200	204.8	128	2P/1P
7713P											1P
7663	56	112	2.00	3.50	240	256	8	3200	204.8	128	2P/1P
7643	48	96	2.30	3.60	225	256	8	3200	204.8	128	2P/1P
7543	32	64	2.80	3.70	225	256	8	3200	204.8	128	2P/1P
7543P											1P
7513	32	64	2.60	3.65	200	128	8	3200	204.8	128	2P/1P
75F3	32	64	2.95	4.00	280	256	8	3200	204.8	128	2P/1P
74F3	24	48	3.20	4.00	240	256	8	3200	204.8	128	2P/1P
73F3	16	32	3.50	4.00	240	256	8	3200	204.8	128	2P/1P
72F3	8	16	3.70	4.10	180	256	8	3200	204.8	128	2P/1P

Intel Xeon "Ice Lake"

Xeon CPUs in "Whitley" Servers and Workstations



Continued specialization of CPU SKUs



FOUR & EIGHT SOCKET SCALABLE PERFORMANCE

SKU	CORES	BASE (GHz)	SINGLE CORE TURBO (GHz)	ALL CORE TURBO (GHz)	CACHE (MB)	TDP (Watts)	Support for Intel Optane Persistent Memory 200 Series	Recommended Customer Pricing (RCP) in $ US Dollars
8380HL	28	2.9	4.3	3.8	38.5	250	Yes	$13012
8380H	28	2.9	4.3	3.8	38.5	250	Yes	$10009
8376HL	28	2.6	4.3	3.5	38.5	205	Yes	$11772
8376H	28	2.6	4.3	3.5	38.5	205	Yes	$8719
8360HL	24	3.0	4.2	3.8	33	225	Yes	$7203
8360H	24	3.0	4.2	3.8	33	225	Yes	$4200
8356H	8	3.9	4.4	4.3	35.75	190	Yes	$3400
8354H	18	3.1	4.3	4.0	24.75	205	Yes	$3500
8353H	18	2.5	3.8	3.3	24.75	150	Yes	$3003
6348H	24	2.3	4.2	3.1	33	165	Yes	$2700
6330H	24	2.0	3.7	2.8	33	150	Yes	$1894
6328HL	16	2.8	4.3	3.7	22	165	Yes	$4779
6328H	16	2.8	4.3	3.7	22	165	Yes	$1776
5320H	20	2.4	4.2	3.3	27.5	150	Yes	$1555
5318H	18	2.5	3.8	3.3	24.75	150	Yes	$1273

H and HL SKUs are only supported on a unique 4 or 8-socket platform. Please contact your hardware provider for a list of system availability supporting your specific SKU configuration.

H SKUs are configured to support up to 1.12 TB of system memory, per processor.
HL SKUs are configured to support up to 4.5 TB of system memory, per processor.
H and HL SKUs are validated for up to 256 GB capacity DRAM memory modules, as of March 2021.

H and HL SKUs support Intel Optane persistent memory 200 series in a 4-socket platform only.
H SKUs are validated for up to 768 GB of Intel Optane persistent memory 200 series, per processor.
HL SKUs are validated for up to 3 TB of Intel Optane persistent memory 200 series, per processor.

6330H, 6328H, 6328HL & 5320H processors include Intel Speed Select technology (Intel SST) supporting Intel SST Core Power (SST-CP) and Intel SST Turbo Frequency (SST-TF) capabilities.

3rd Gen Intel Xeon Scalable Processors

intel.com/xeonscalable

Intel may make changes to specifications and product descriptions at any time, without notice.
Please visit intel.com/xeon or contact your Intel representative to obtain the latest product specifications.

Intel processor numbers are not a measure of performance. Processor numbers differentiate features within each processor family, not across different processor families. All processors support Intel Virtualization Technology (Intel VT-x). No product or component can be absolutely secure.

OPTIMIZED FOR HIGHEST PER-CORE SCALABLE PERFORMANCE

SKU	CORES	BASE (GHz)	SINGLE CORE TURBO (GHz)	ALL CORE TURBO (GHz)	CACHE (MB)	TDP (Watts)	Support for Intel Optane Persistent Memory 200 Series	Intel SGX Enclave Capacity (Per Processor)	Recommended Customer Pricing (RCP) in $ US Dollars
8380	40	2.3	3.4	3.0	60	270	Yes	512 GB	$8099
8368	38	2.4	3.4	3.2	57	270	Yes	512 GB	$6302
8362	32	2.8	3.6	3.5	48	265	Yes	64 GB	$5448
8360Y	36	2.4	3.5	3.1	54	250	Yes	64 GB	$4702
8358	32	2.6	3.4	3.3	48	250	Yes	64 GB	$3950
6348	28	2.6	3.5	3.4	42	235	Yes	64 GB	$3072
6342	24	2.8	3.5	3.3	36	230	Yes	64 GB	$2529
6354	18	3.0	3.6	3.6	39	205	Yes	64 GB	$2445
6346	16	3.1	3.6	3.6	36	205	Yes	64 GB	$2300
6334	8	3.6	3.7	3.6	18	165	Yes	64 GB	$2214
6326	16	2.9	3.5	3.3	24	185	Yes	64 GB	$1300
5317	12	3.0	3.6	3.4	18	150	Yes	64 GB	$950
5315Y	8	3.2	3.6	3.5	12	140	Yes	64 GB	$895

SCALABLE PERFORMANCE

8352Y	32	2.2	3.4	2.8	48	205	Yes	64 GB	$3450
6338	32	2.0	3.2	2.6	48	205	Yes	64 GB	$2612
6336Y	24	2.4	3.6	3.0	36	185	Yes	64 GB	$1977
6330	28	2.0	3.1	2.6	42	205	Yes	64 GB	$1894
5320	26	2.2	3.4	2.8	39	185	Yes	64 GB	$1555
5318Y	24	2.1	3.4	2.6	36	165	Yes	64 GB	$1273
4316	20	2.3	3.4	2.8	30	150		8 GB	$1002
4314	16	2.4	3.4	2.9	24	135	Yes	8 GB	$694
4310	12	2.1	3.3	2.7	18	120		8 GB	$501
4309Y	8	2.8	3.6	3.4	12	105		8 GB	$501

Y Supports Intel Speed Select Technology – Performance Profile 2.0 (Intel SST-PP)

All 8300, 6300, 5300 and 4300 processors, Non-H/HL SKUs, are supported on a unique 1 or 2 socket platform. Please contact your hardware provider for a list of system availability supporting your specific SKU configuration.

All 8300, 6300, 5300 and 4300 processors, Non-H/HL SKUs, are configured to support up to 6 TB of system memory, per processor. Intel has validated for up to 4 TB of Intel Optane persistent memory 200 series, per processor. Intel has validated for up to 256 GB capacity DRAM memory modules, as of March 2021.

Unless noted, all 8300, 6300 and 5300 processors, Non-H/HL SKUs, include support for Intel Speed Select technology (Intel SST) featuring Intel SST Base Frequency (SST-BF), Intel SST Core Power (SST-CP) and Intel SST Turbo Frequency (SST-TF) capabilities.

M, P, Q, V SKUs and 8362 do not include support Intel Speed Select Technology Base Frequency (SST-BF).

SKUs SUPPORTING MAXIMUM INTEL SGX ENCLAVE CAPACITY

SKU	CORES	BASE (GHz)	SINGLE CORE TURBO (GHz)	ALL CORE TURBO (GHz)	CACHE (MB)	TDP (Watts)	Support for Intel Optane Persistent Memory 200 Series	Intel SGX Enclave Capacity (Per Processor)	Recommended Customer Pricing (RCP) in $ US Dollars
8380	40	2.3	3.4	3.0	60	270	Yes	512 GB	$8099
8368Q	38	2.6	3.7	3.3	57	270	Yes	512 GB	$6743
8368	38	2.4	3.4	3.2	57	270	Yes	512 GB	$6302
8352S	32	2.2	3.4	2.8	48	205	Yes	512 GB	$4046
5318S	24	2.1	3.4	2.6	36	165	Yes	512 GB	$1667

8352S and 5318S support Intel Speed Select Technology – Performance Profile 2.0 (Intel SST-PP)

CLOUD OPTIMIZED FOR VM UTILIZATION

8358P	32	2.6	3.4	3.2	48	240	Yes	8 GB	$3950
8352V	36	2.1	3.5	2.5	54	195	Yes	8 GB	$3450

P IaaS Cloud Specialized Processor, **V** SaaS Cloud Specialized Processor
8352V supports Intel Speed Select Technology – Performance Profile 2.0 (Intel SST-PP)

LIQUID COOLED

8368Q	38	2.6	3.7	3.3	57	270	Yes	512 GB	$6743

8368Q supports up to 512 GB Intel Software Guard Extensions (Intel SGX) enclave capacity

NETWORKING/NFV OPTIMIZED

8351N	36	2.4	3.5	3.1	54	225	Yes	64 GB	$3027
6338N	32	2.2	3.5	2.7	48	185	Yes	64 GB	$2795
6330N	28	2.2	3.4	2.6	42	165	Yes	64 GB	$2029
5318N	24	2.1	3.4	2.7	36	150	Yes	64 GB	$1375

8351N is supported in a one-socket configuration only
5318N supports Intel Speed Select Technology – Performance Profile 2.0 (Intel SST-PP)

MEDIA PROCESSING OPTIMIZED

8352M	32	2.3	3.5	2.8	48	185	Yes	64 GB	$3864

Optimized for processing AI and media workloads and services.

LONG-LIFE USE AND NEBS-THERMAL FRIENDLY

6338T	24	2.1	3.4	2.7	36	165	Yes	64 GB	$2742
5320T	20	2.3	3.5	2.9	30	150	Yes	64 GB	$1727
4310T	10	2.3	3.4	2.9	15	105		8 GB	$555

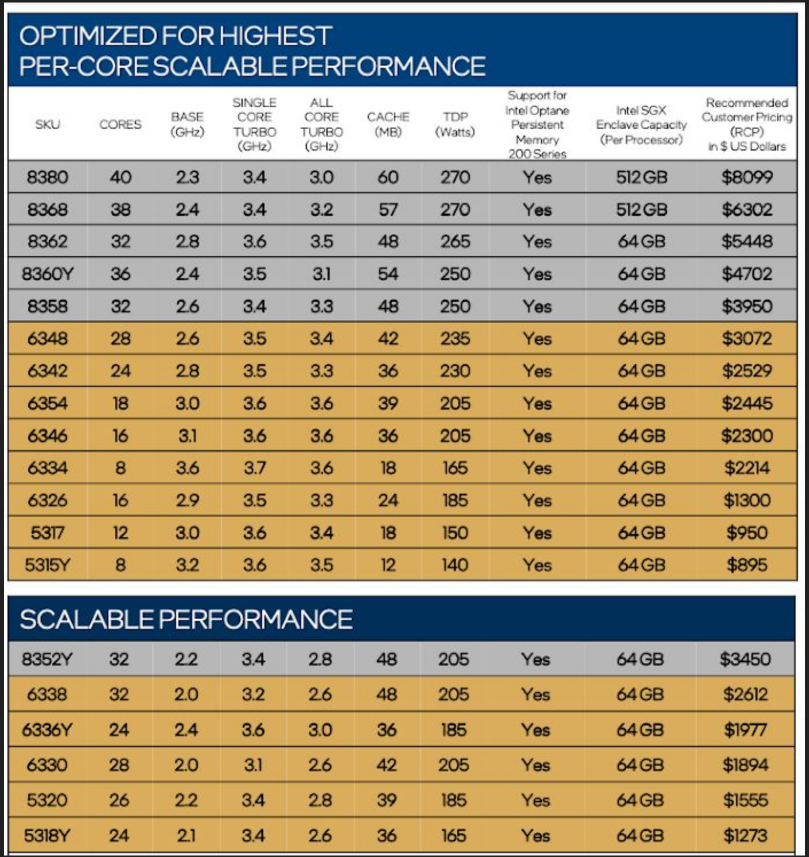
Support for up to 10-year reliability, higher Tcase.

SINGLE-SOCKET OPTIMIZED

8351N	36	2.4	3.5	3.1	54	225	Yes	64 GB	$3027
6314U	32	2.3	3.4	2.9	48	205	Yes	64 GB	$2600
6312U	24	2.4	3.6	3.1	36	185	Yes	64 GB	$1450

Supported in one-socket configurations only

High core counts; High wattages

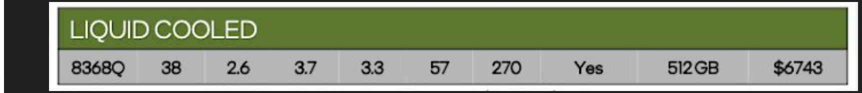


OPTIMIZED FOR HIGHEST PER-CORE SCALABLE PERFORMANCE

SKU	CORES	BASE (GHz)	SINGLE CORE TURBO (GHz)	ALL CORE TURBO (GHz)	CACHE (MB)	TDP (Watts)	Support for Intel Optane Persistent Memory 200 Series	Intel SGX Enclave Capacity (Per Processor)	Recommended Customer Pricing (RCP) in $ US Dollars
8380	40	2.3	3.4	3.0	60	270	Yes	512 GB	$8099
8368	38	2.4	3.4	3.2	57	270	Yes	512 GB	$6302
8362	32	2.8	3.6	3.5	48	265	Yes	64 GB	$5448
8360Y	36	2.4	3.5	3.1	54	250	Yes	64 GB	$4702
8358	32	2.6	3.4	3.3	48	250	Yes	64 GB	$3950
6348	28	2.6	3.5	3.4	42	235	Yes	64 GB	$3072
6342	24	2.8	3.5	3.3	36	230	Yes	64 GB	$2529
6354	18	3.0	3.6	3.6	39	205	Yes	64 GB	$2445
6346	16	3.1	3.6	3.6	36	205	Yes	64 GB	$2300
6334	8	3.6	3.7	3.6	18	165	Yes	64 GB	$2214
6326	16	2.9	3.5	3.3	24	185	Yes	64 GB	$1300
5317	12	3.0	3.6	3.4	18	150	Yes	64 GB	$950
5315Y	8	3.2	3.6	3.5	12	140	Yes	64 GB	$895

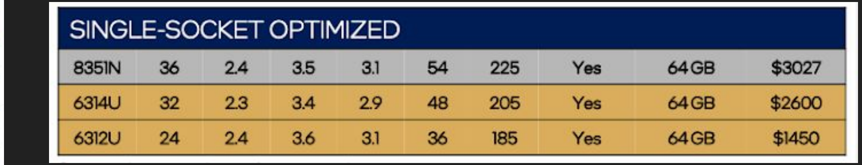
SCALABLE PERFORMANCE

8352Y	32	2.2	3.4	2.8	48	205	Yes	64 GB	$3450
6338	32	2.0	3.2	2.6	48	205	Yes	64 GB	$2612
6336Y	24	2.4	3.6	3.0	36	185	Yes	64 GB	$1977
6330	28	2.0	3.1	2.6	42	205	Yes	64 GB	$1894
5320	26	2.2	3.4	2.8	39	185	Yes	64 GB	$1555
5318Y	24	2.1	3.4	2.6	36	165	Yes	64 GB	$1273



LIQUID COOLED

8368Q	38	2.6	3.7	3.3	57	270	Yes	512 GB	$6743



SINGLE-SOCKET OPTIMIZED

8351N	36	2.4	3.5	3.1	54	225	Yes	64 GB	$3027
6314U	32	2.3	3.4	2.9	48	205	Yes	64 GB	$2600
6312U	24	2.4	3.6	3.1	36	185	Yes	64 GB	$1450

NVIDIA GTC 2021

GTC Digital: the whole conference is online!

- HUNDREDS of recorded talks and panels
- PDF downloads of most presentations
- Plus self-paced training, demos, podcasts, and "Connect with Experts" sessions

https://gtc21.event.nvidia.com/

https://www.nvidia.com/en-us/gtc/on-demand/

Nanoseconds per Day

Simulating COVID on the world's 2nd-largest supercomputer

~305 million atoms in COVID virus

https://gtc21.event.nvidia.com/media/1_guble5xm

Nanoseconds per Day

Simulating COVID on the world's 2nd-largest supercomputer

SUMMIT capable of 64~128 ns/day

https://gtc21.event.nvidia.com/media/1_guble5xm

Nanoseconds per Day

Simulating COVID on the world's 2nd-largest supercomputer

That works out to:

0.000000128th of one second

https://gtc21.event.nvidia.com/media/1_guble5xm

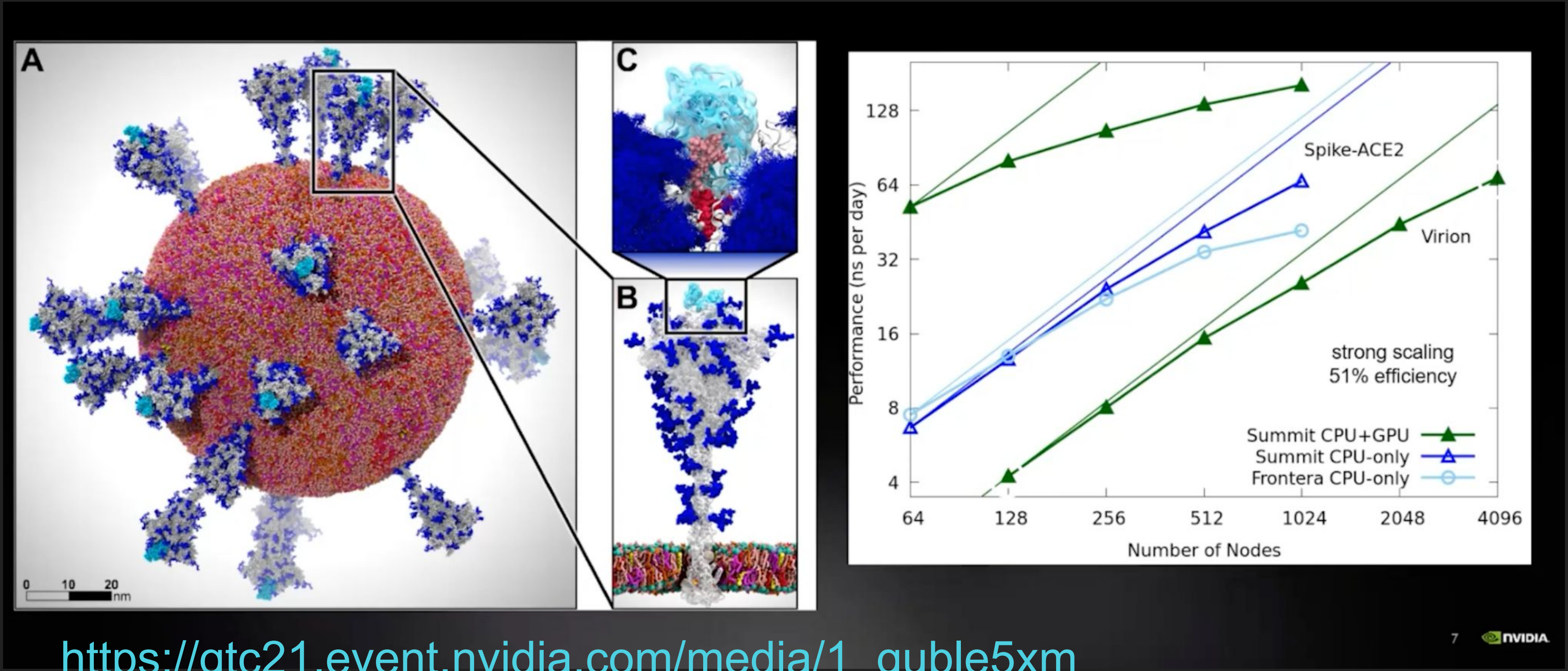
Nanoseconds per Day

Simulating COVID on the world's 2nd-largest supercomputer

https://gtc21.event.nvidia.com/media/1_guble5xm

But the process of a virion entering a cell takes minutes

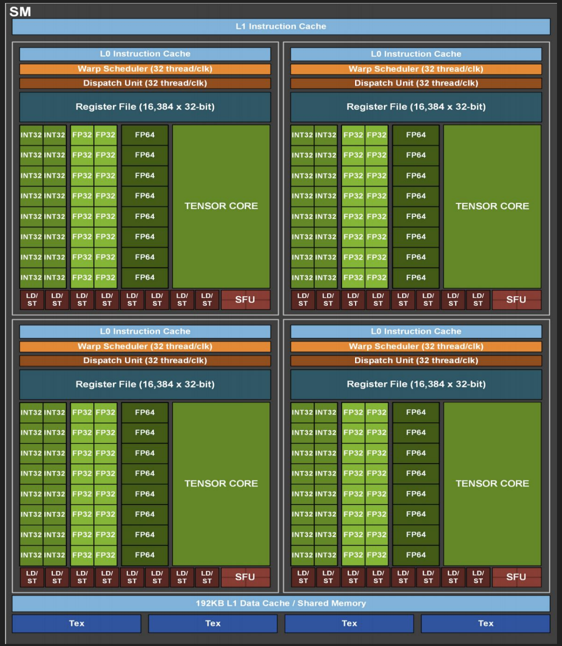
Convergence of AI and HPC to Solve Grand Challenges



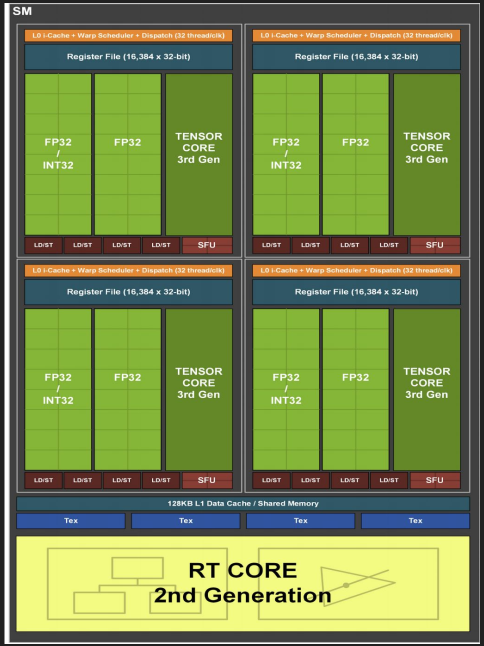
https://gtc21.event.nvidia.com/media/1_guble5xm

New NVIDIA server GPUs

Two paths for NVIDIA GPUs (not to scale)

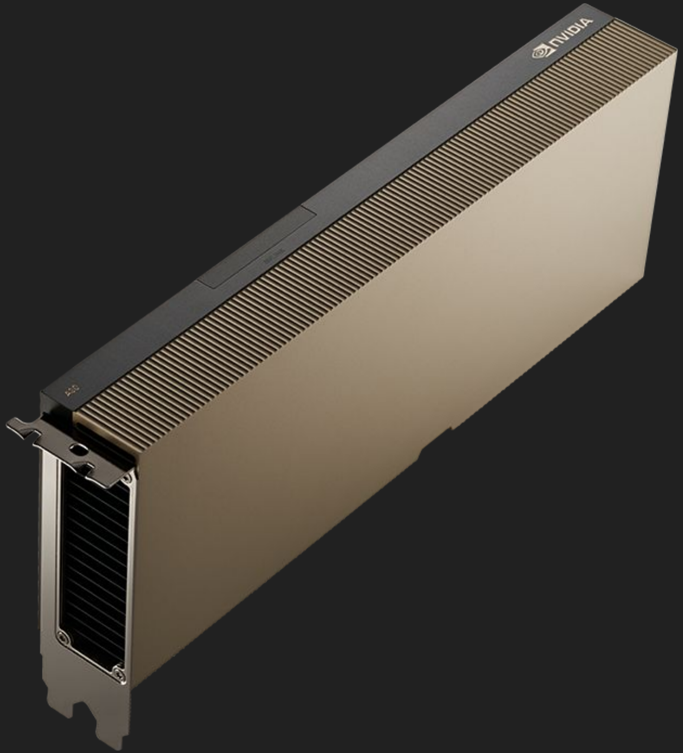

SM
L1 Instruction Cache
L0 Instruction Cache
Warp Scheduler (32 thread/clk)
Dispatch Unit (32 thread/clk)
Register File (16,384 x 32-bit)
INT32
FP32
FP64
TENSOR CORE
LD/ST
SFU
192KB L1 Data Cache / Shared Memory
Tex

NVIDIA GA100 Architecture
(just for DGX and servers)

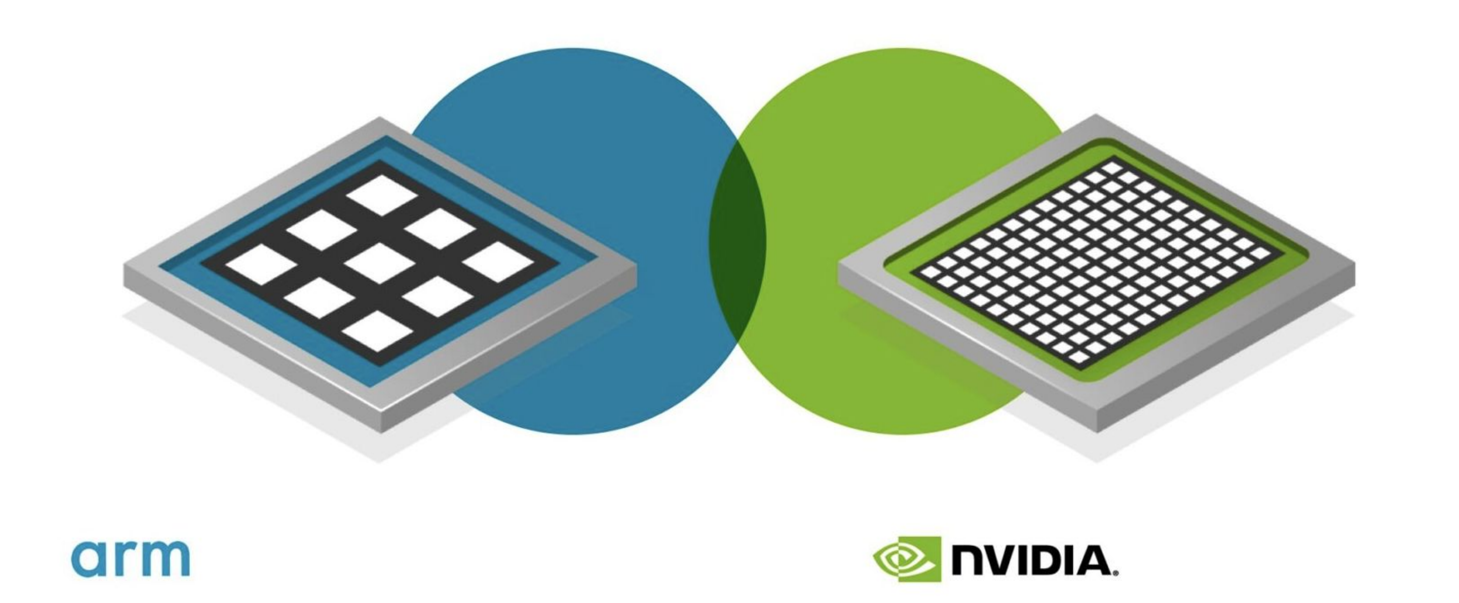

SM
L0 i-Cache + Warp Scheduler + Dispatch (32 thread/clk)
Register File (16,384 x 32-bit)
FP32 / INT32
FP32
TENSOR CORE 3rd Gen
LD/ST
SFU
128KB L1 Data Cache / Shared Memory
Tex
RT CORE 2nd Generation

NVIDIA GA102 Architecture
(workstations and servers)

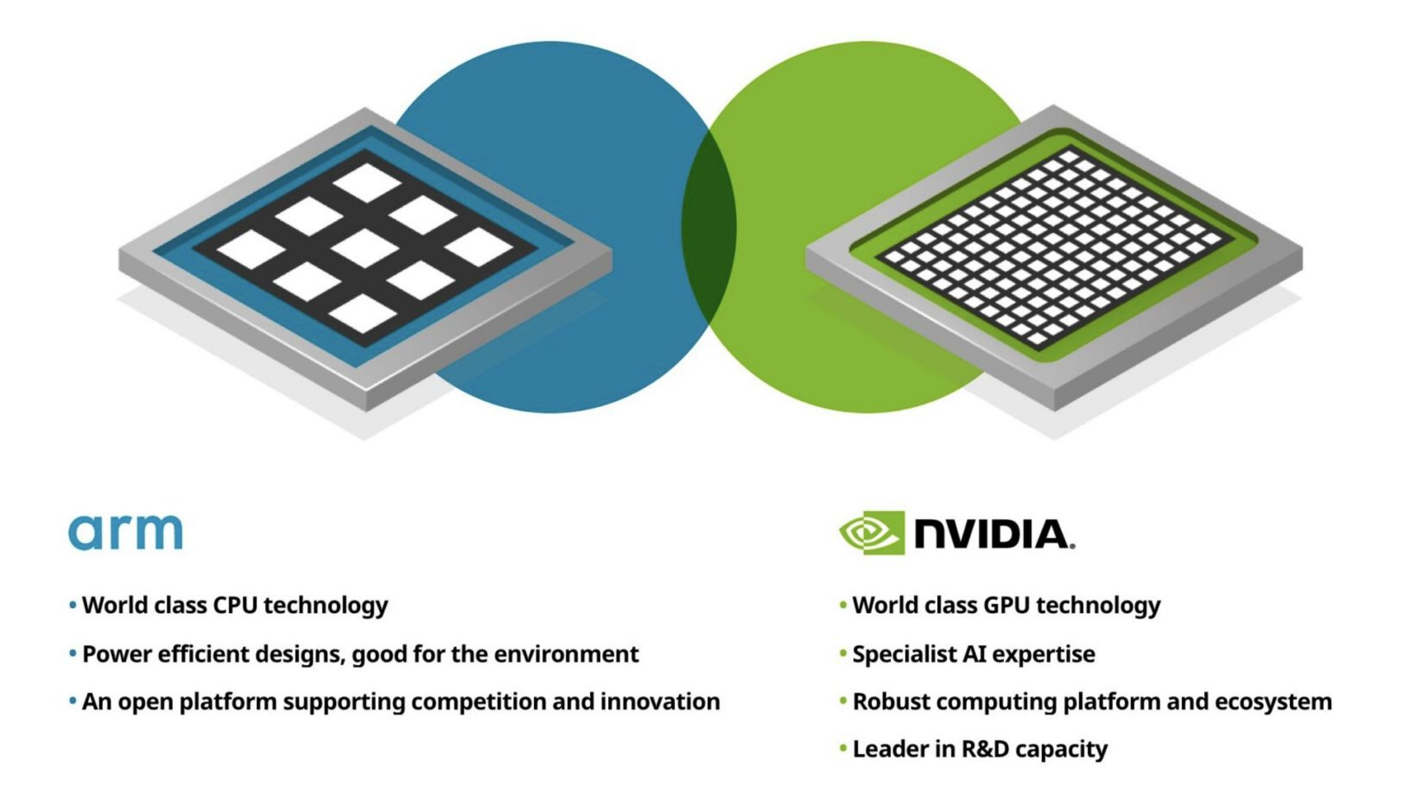
A30 and A10 GPUs (with an A16 to come)



ARM + NVIDIA



What does Arm + NVIDIA offer the world?



arm

- World class CPU technology
- Power efficient designs, good for the environment
- An open platform supporting competition and innovation

NVIDIA.

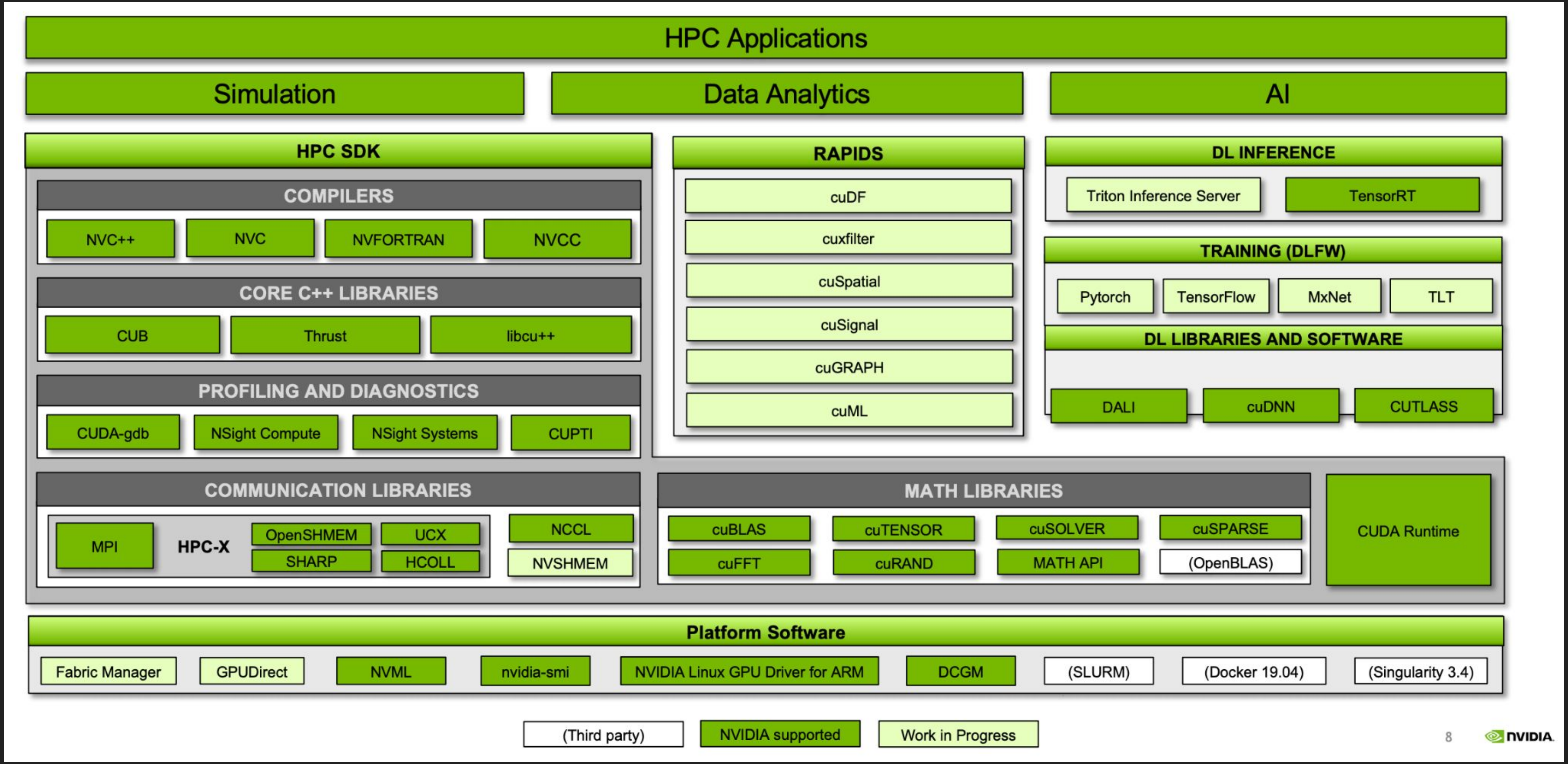
- World class GPU technology
- Specialist AI expertise
- Robust computing platform and ecosystem
- Leader in R&D capacity

NVIDIA ARM HPC Dev Kit

- Delivers a validated system for quick and easy bring-up in familiar HPC environments
- Provides a stable hardware and software platform for development and performance analysis of accelerated HPC, AI, and scientific computing applications
- Enables experimentation and characterization of high-performance, NVIDIA-accelerated, Arm server-based system architectures

Model	Gigabyte G242-P32, 2U server
CPU	1x Ampere Altra Q80-30
Memory	512GB DDR4
Storage	6TB SAS/SATA 3.5”
GPU	2x A100 PCIE 40GB
Network	NVIDIA® BlueField®-2 E-Series DPU 200GbE/HDR, PCIe Gen4 x16, 16GB on-board DDR
Power	1600W
Availability	**July 2021**

NVIDIA ARM HPC Software



HPC Applications

Simulation

Data Analytics

AI

HPC SDK

COMPILERS

NVC++

NVC

NVFORTRAN

NVCC

CORE C++ LIBRARIES

CUB

Thrust

libcu++

PROFILING AND DIAGNOSTICS

CUDA-gdb

NSight Compute

NSight Systems

CUPTI

COMMUNICATION LIBRARIES

MPI

HPC-X

OpenSHMEM

UCX

SHARP

HCOLL

NCCL

NVSHMEM

MATH LIBRARIES

cuBLAS

cuTENSOR

cuSOLVER

cuSPARSE

cuFFT

cuRAND

MATH API

(OpenBLAS)

CUDA Runtime

RAPIDS

cuDF

cuxfilter

cuSpatial

cuSignal

cuGRAPH

cuML

DL INFERENCE

Triton Inference Server

TensorRT

TRAINING (DLFW)

Pytorch

TensorFlow

MxNet

TLT

DL LIBRARIES AND SOFTWARE

DALI

cuDNN

CUTLASS

Platform Software

Fabric Manager

GPUDirect

NVML

nvidia-smi

NVIDIA Linux GPU Driver for ARM

DCGM

(SLURM)

(Docker 19.04)

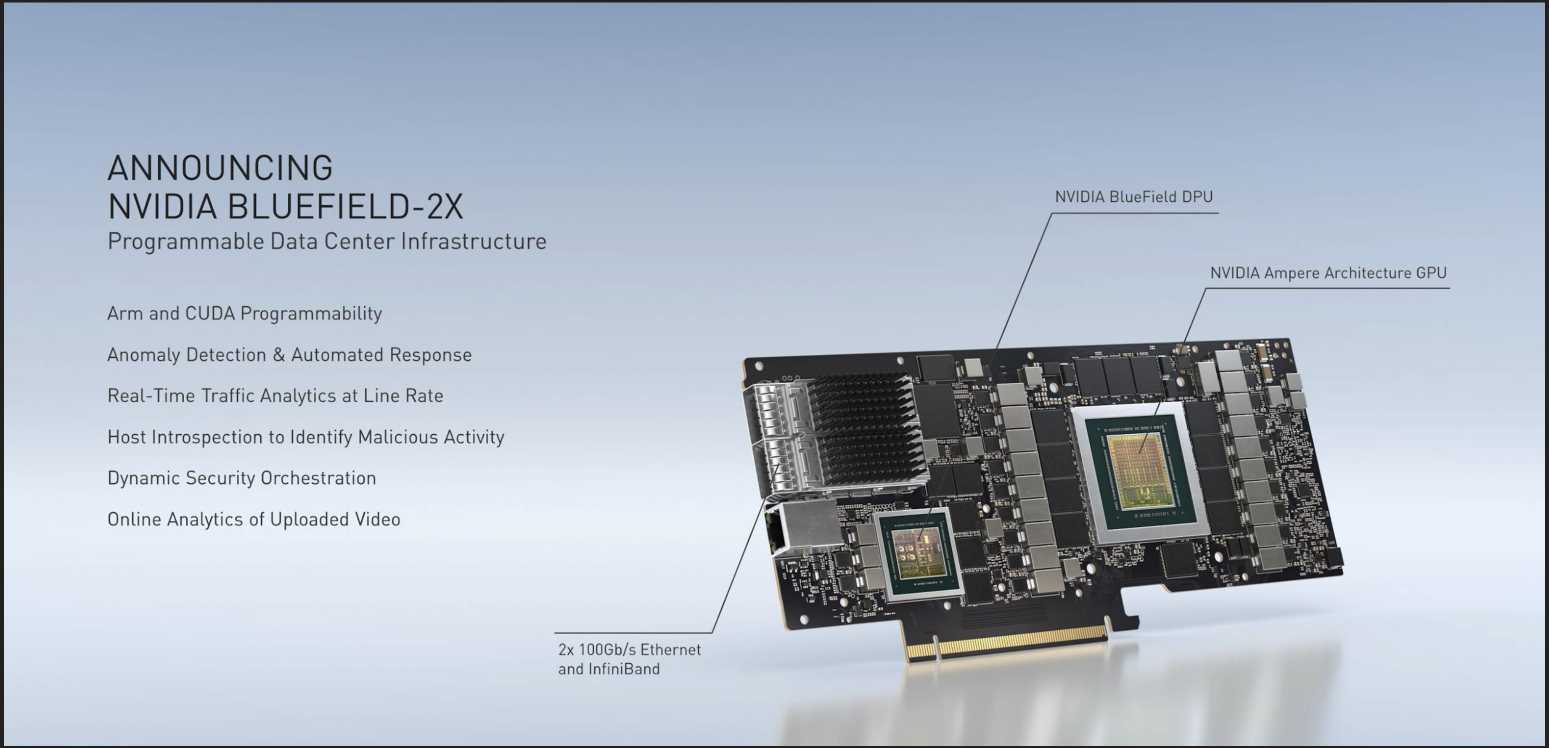
(Singularity 3.4)

(Third party)

NVIDIA supported

Work in Progress

NVIDIA Bluefield DPU - as announced in 2020



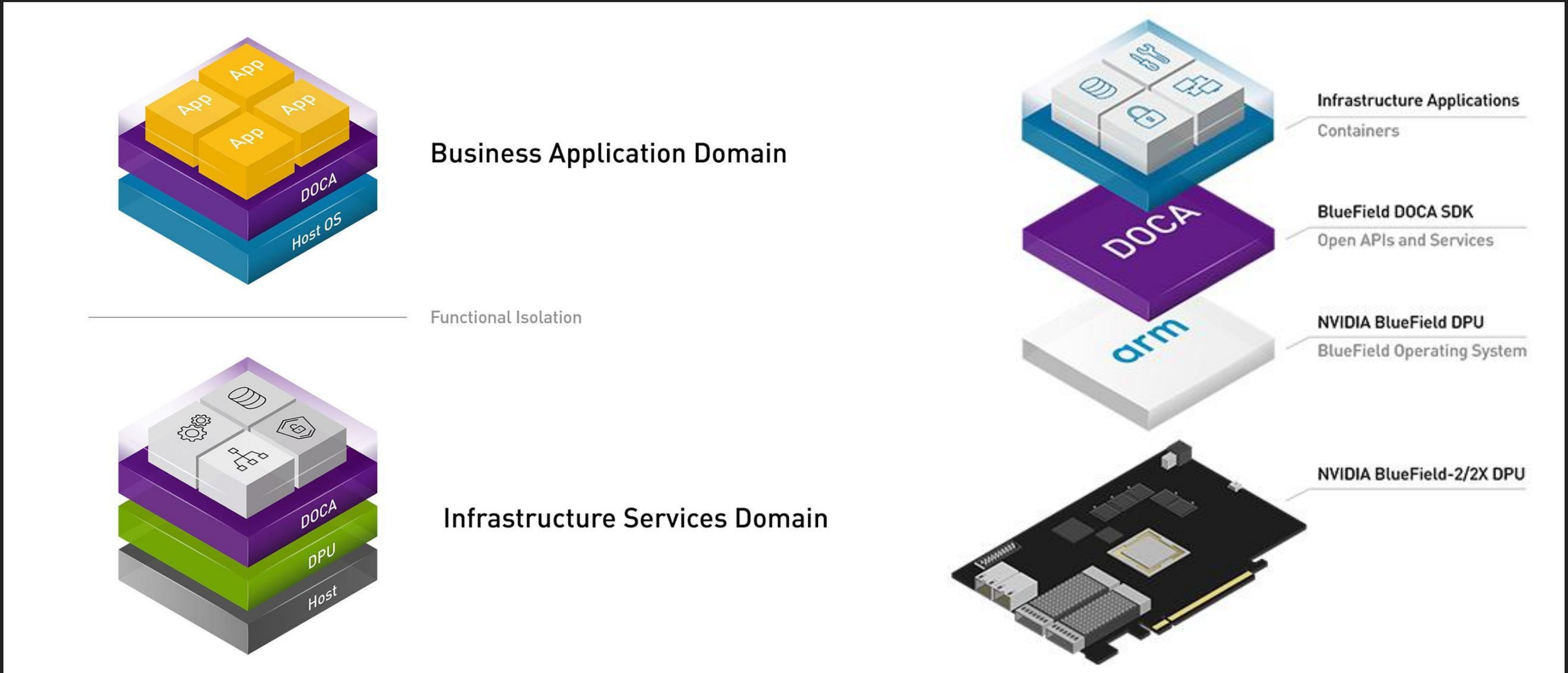
ANNOUNCING
NVIDIA BLUEFIELD-2X

Programmable Data Center Infrastructure

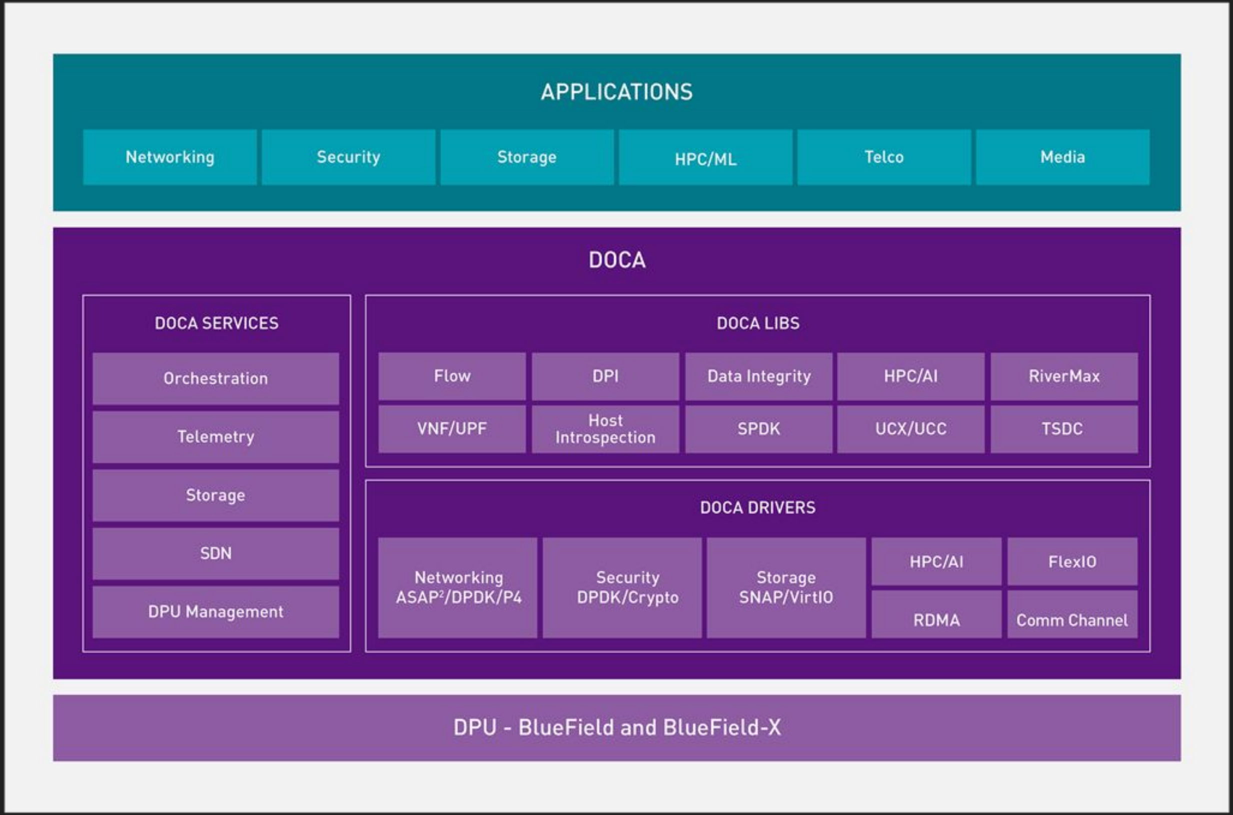
- Arm and CUDA Programmability
- Anomaly Detection & Automated Response
- Real-Time Traffic Analytics at Line Rate
- Host Introspection to Identify Malicious Activity
- Dynamic Security Orchestration
- Online Analytics of Uploaded Video

NVIDIA BlueField DPU
NVIDIA Ampere Architecture GPU
2x 100Gb/s Ethernet and InfiniBand

Data Center Infrastructure-on-a-Chip Architecture


App
App
App
App
DOCA
Host OS
Business Application Domain
Functional Isolation
DOCA
DPU
Host
Infrastructure Services Domain
Infrastructure Applications
Containers
DOCA
BlueField DOCA SDK
Open APIs and Services
arm
NVIDIA BlueField DPU
BlueField Operating System
NVIDIA BlueField-2/2X DPU

NVIDIA DOCA



APPLICATIONS

Networking | Security | Storage | HPC/ML | Telco | Media

DOCA

DOCA SERVICES

- Orchestration
- Telemetry
- Storage
- SDN
- DPU Management

DOCA LIBS

- Flow
- DPI
- Data Integrity
- HPC/AI
- RiverMax
- VNF/UPF
- Host Introspection
- SPDK
- UCX/UCC
- TSDC

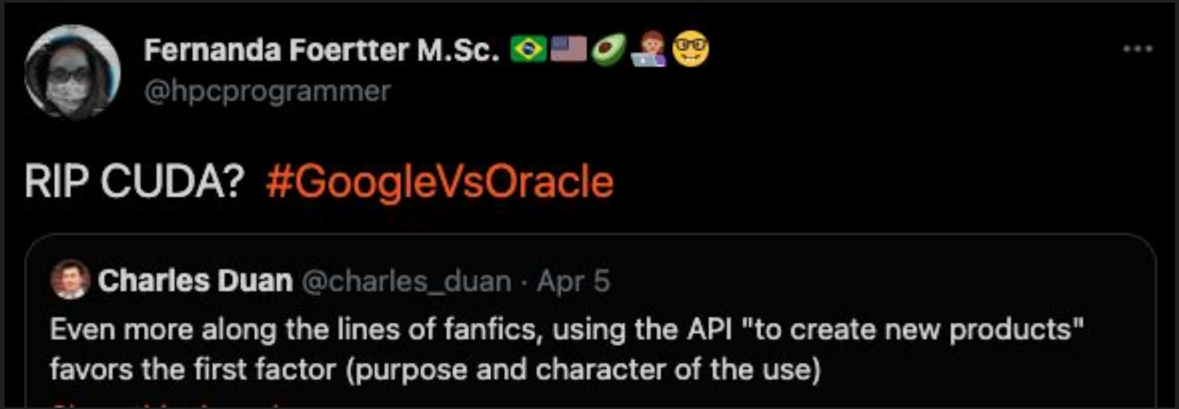
DOCA DRIVERS

- Networking ASAP^2/DPDK/P4
- Security DPDK/Crypto
- Storage SNAP/VirtIO
- HPC/AI
- FlexIO
- RDMA
- Comm Channel

DPU - BlueField and BlueField-X

Oracle vs Google - outlook without API copyrights

I am not a lawyer, but might we see CUDA implementations from other vendors?



Fernanda Foertter M.Sc. 🇧🇷🇺🇸🥑👩🏽‍💻🤓
@hpcprogrammer

RIP CUDA? #GoogleVsOracle

Charles Duan @charles_duan · Apr 5
Even more along the lines of fanfics, using the API "to create new products" favors the first factor (purpose and character of the use)