THE INTERNATIONAL MEETING
CONFERENCES | WORKSHOPS | EXHIBITION

22-23-24 JUNE
DIGITAL EVENT

Forum Teratec 2021

Unlock the future !

SIMULATION |
HPC | HPDA
AI | QUANTUM

PLATINUM SPONSORS

GOLD SPONSORS

SILVER SPONSORS

PARTENAIRE EUROPA VILLAGE

arm

SW defined cars: HPC, from the cloud to the dashboard for an amazing driver experience

Deployment strategy for true self-driving vehicles

Eric.lalardie@arm.com

+33 6 07 83 09 60

24 June 2021

Agenda

- Intro
- **The Changing Automotive Development and Deployment Paradigm**
  - Enabling compelling autonomous system solutions
- Arm in DC/HPC
  - Enabling cloud to edge development/deployments (CI/CD)
- Conclusion
- Neoverse V1 introduction
- Neoverse N2 Introduction

Sparking the World's Potential

arm

Semiconductor IP Business

The global leader in the development of licensable compute technology

- R&D outsourcing for semiconductor companies

Focused on freedom and flexibility to innovate

- Technology reused across multiple applications

With a partnership based culture & business model

- Licensees take advantage of learnings from a uniquely collaborative ecosystem

1,910
licenses, growing by 100+ every year

530 licensees
Industry leaders and high-growth start-ups; chip companies and OEMs

180+bn
Arm-based chips shipped to-date

23.7bn
Arm-based chips shipped in 2020

The Changing Automotive Development and Deployment Paradigm

Evolving Automotive Compute Architecture

Evolution of Software Defined Systems

Cloud-Native In Automotive

Some of the Challenges the SW-defined Car developers are facing

Software Portability
Write Once, Run Everywhere

+

Cloud-Native Mixed Criticality
Write for the Cloud, Run in the Car

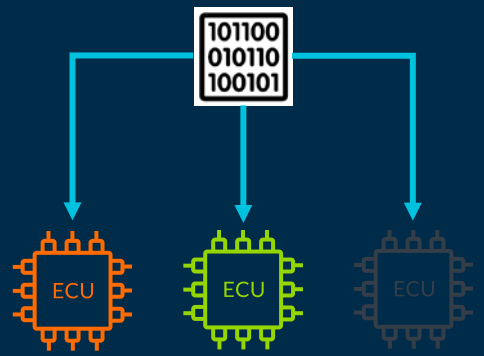
+

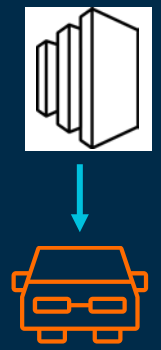
Seamless Deployability
Execution Environment Parity

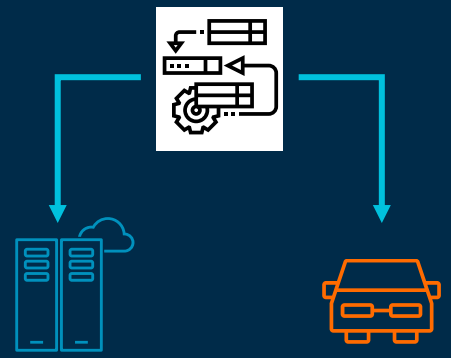
=

The Software-Defined Vehicle

Need For Standards Through The Ecosystem







Requiring a Seamless Continuum From Development to Deployment

Key Automotive trends

Application consolidation

Software-defined functionality

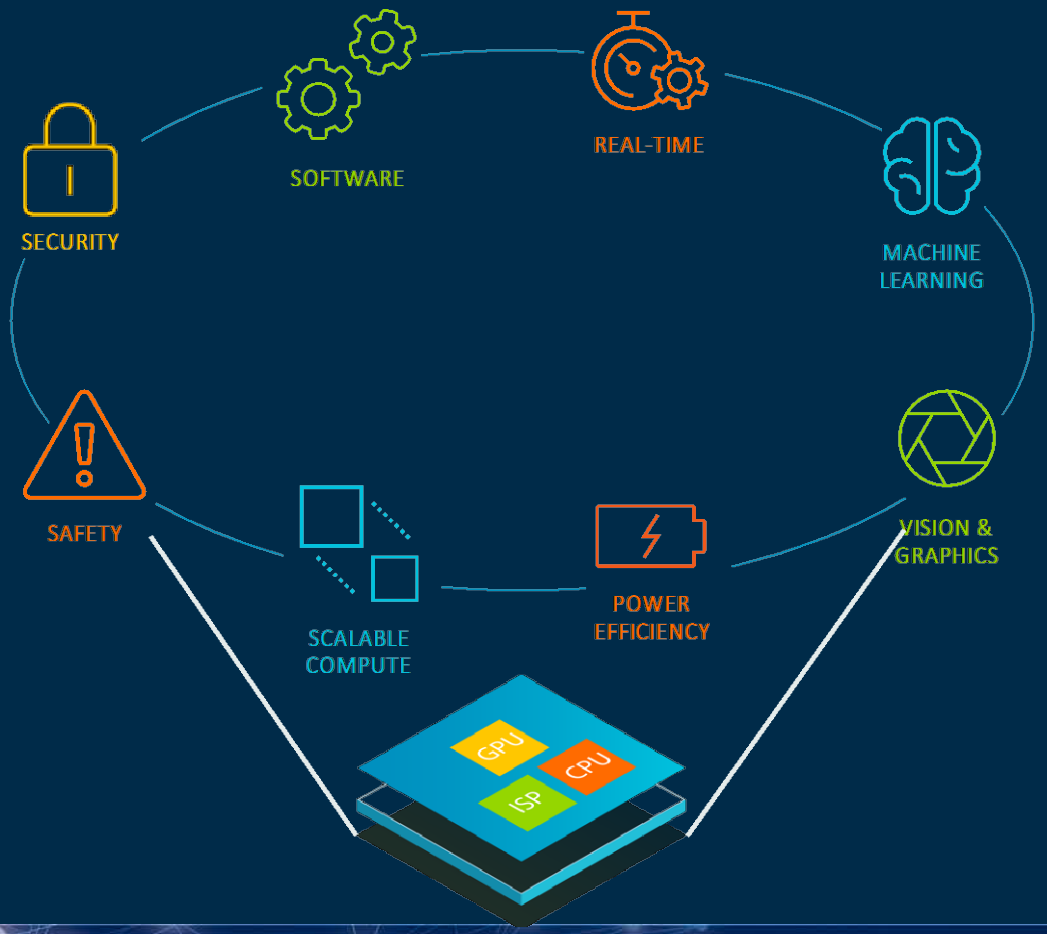
High performance compute

Energy efficiency

Richer user experiences

Functional Safety + Security

Key Arm capabilities



Dedicated investment in Automotive segment

Underpinned by IP developed ground-up for functional safety

Segment Focus

- Digital Cockpit
- ADAS Autonomous
- Vehicle Motion & Dynamics

Key IP Attributes

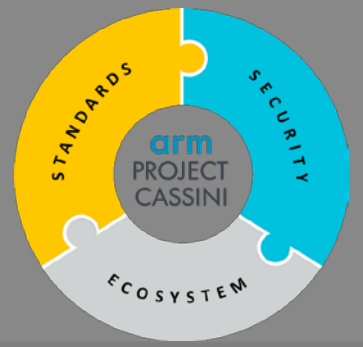
- Functional Safety
- Scalable Processing
- Machine Learning
- Heterogenous Compute
- Reusable Software Stack

Ecosystem Initiatives

arm SystemReady CERTIFIED V2.0

AVCC AUTONOMOUS VEHICLE COMPUTING CONSORTIUM

Holistic Security for the Edge

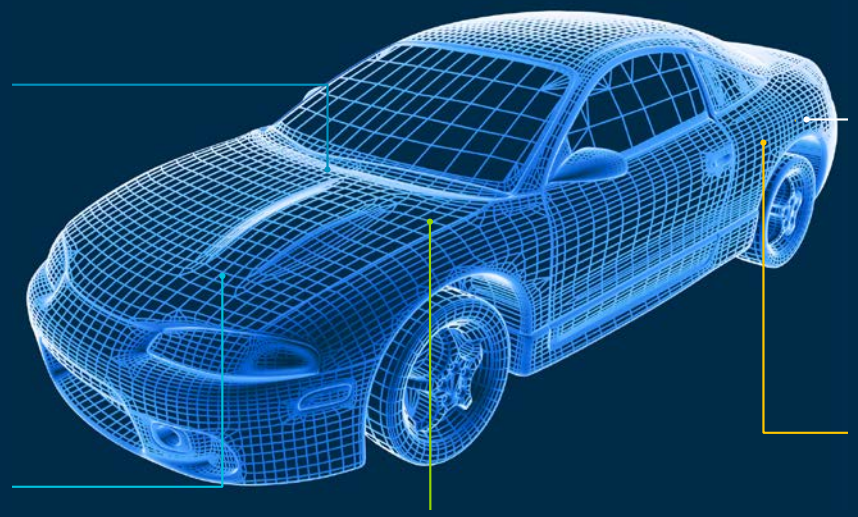


psacertified™

The Autonomous Vehicle Computing Consortium (AVCC)

- AVCC is a group of automotive and technology industry leaders coming together
- AVCC aim to help accelerate mass production of safe and affordable vehicles with automated and assisted driving solutions,
  - as defined by the SAE practice J3016 levels 1-5.
- All automotive and technology companies are encouraged and welcomed to join the AVCC.

Arm solutions span the full range of automotive systems

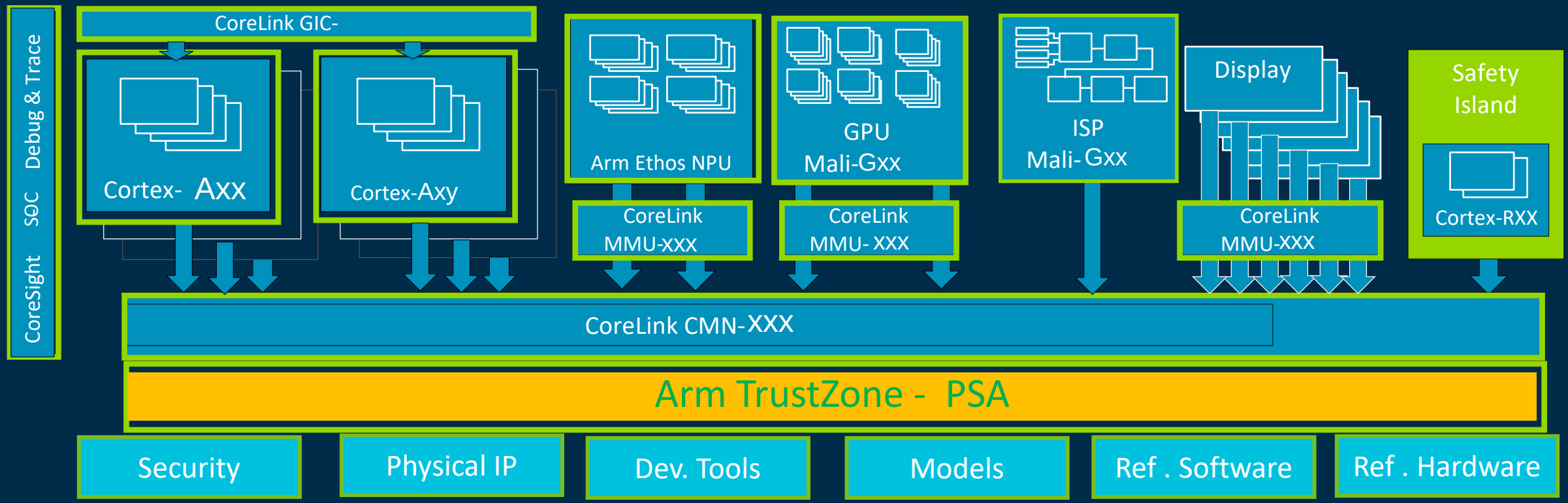


Continuously supporting the automotive industry since **1996**

15
Top automotive chip makers license Arm IP

>60%
Share of all Infotainment & ADAS application processors

Comprehensive range of technologies and tools
Example central compute sub-system



arm NEOVERSE

Shaping the future of Cloud, Infrastructure, HPC

Enabling native edge to cloud development & deployment

Hyperscale & Cloud Computing

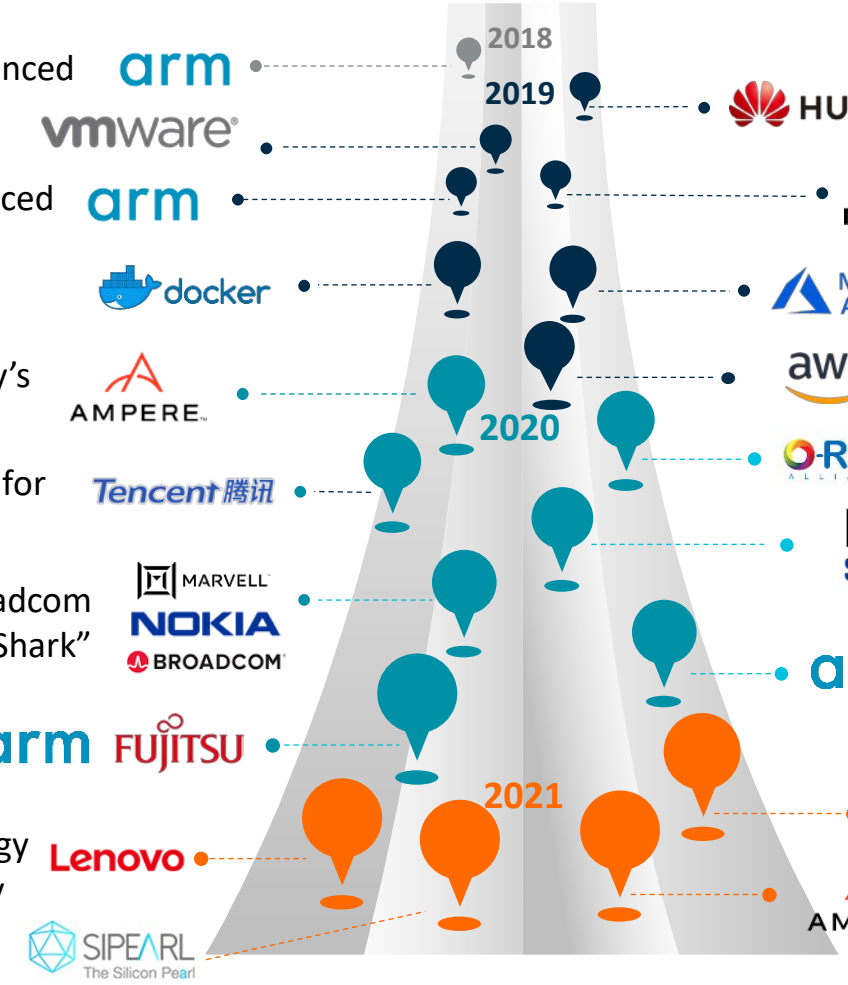
HPC

5G

Infrastructure Edge

arm PROJECT CASSINI

Arm Neoverse Momentum



- 2018 — Arm Neoverse Announced (arm)
- 2019 — VMware demonstrates ESXi on ARM (vmware)
- 2019 — Neoverse N1 & E1 Platform announced (arm)
- 2019 — Docker and Arm Partner for Cloud-native Software Dev (docker)
- 2019 — Ampere announced industry's 1st 80-Core Server SoC (AMPERE)
- 2020 — Tencent partners with Huawei for Cloud gaming platform (Tencent 腾讯)
- 2020 — Nokia partner with Marvell and Broadcom on Silicon Tech for 5G "ReefShark" (MARVELL, NOKIA, BROADCOM)
- 2020 — Arm powers the World's Fastest Supercomputer, Fugaku (arm, FUJITSU)
- 2021 — Lenovo adopts Arm technology as it expands 5G its strategy (Lenovo)
- 2021 — Sipearl license Neoverse V1 platform to address EU HPC sovereignty (SIPEARL The Silicon Pearl)
- 2019 — Huawei released Kunpeng 920 CPU and TaiShan server platform (HUAWEI)
- 2019 — Nvidia brought CUDA stack to Arm Platforms (NVIDIA)
- 2019 — Marvell's ThunderX2 Solution for Microsoft Azure Development (Microsoft Azure)
- 2019 — AWS announces Arm-based Graviton2 CPU (aws)
- 2020 — Arm join ORAN Alliance (O-RAN ALLIANCE)
- 2020 — Samsung partner with Marvell and Xilinx on Silicon Technology for 5G (MARVELL, SAMSUNG, XILINX)
- 2020 — Arm announces Neoverse V1 with SVE and Neoverse N2 platforms (arm)
- 2021 — Nvidia announces "Grace" CPU, Bluefield-3 DPU, Developer SDK, Graviton2 GPU plans (NVIDIA)
- 2021 — Oracle Cloud launches Ampere Altra A1 Compute instance family (AMPERE, ORACLE)

Arm Neoverse Platform Roadmap

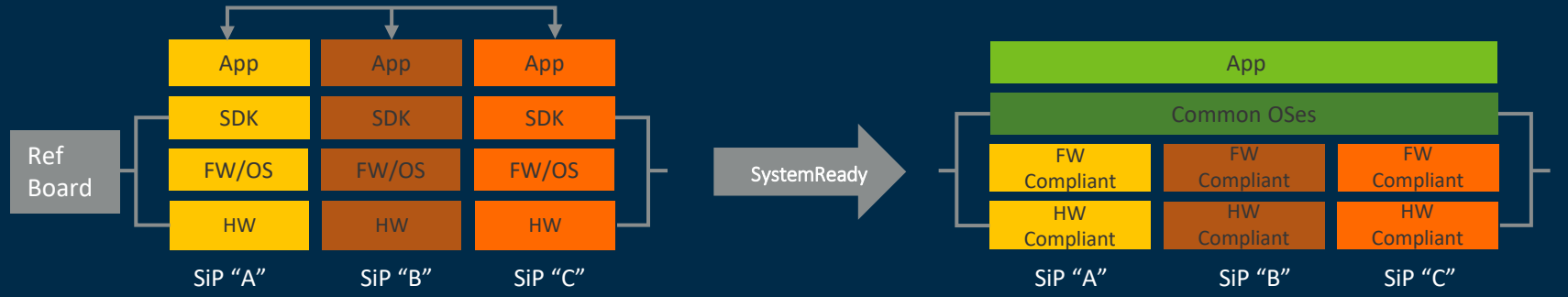
In planning

IP Date	Production	Now		2022+
Platform features	7nm PCIe Gen4, DDR4, HBM2 CCIX 1.0	7/5nm PCIe Gen5, DDR5, HBM2e CCIX 1.1	5nm PCIe Gen5, DDR5, HBM3 CCIX 2.0, CXL 2.0	5/3nm PCIe Gen5/6, DDR5, HBM3 CCIX next, CXL next
V-series  Max ST Perf ML		**V1 Platform (Zeus)** N1+50% ST perf. uplift SVE 2x256b, bFloat16		
N-series  Perf/W Scale out 5G	**N1 Platform** A72+60% ST perf. uplift		**N2 Platform (Perseus)** N1+40% ST perf. uplift SVE2 2x128b, bFloat16	
E-series  Data efficiency	**E1 Platform** A53+2.7x throughput			
				Poseidon Generation Platforms +30% infra WL perf. ML/Vector uplift Greater core density

Arm SystemReady Program Goals

- Promote adoption of standards and tools within the Arm ecosystem to enable open platforms that make it easier for IHV's to embrace Arm as their hardware platform of choice.

AIoT edge ⟷ Cloud edge



- For ODMs/OEMs : Allow them to self-certify and enable faster TTM to market.
- For SiP's: Simplify and accelerate enablement of software and hardware developer ecosystems.
- For OSVs: The OSes would "Just Work" on the certified systems.
- For End Users: Enjoy the "It Just Works" experience.

Broad Ecosystem Support of Cassini and SystemReady

ADVANTECH · altran (Part of Capgemini) · PRO · BROADCOM · Compulab · EUROTECH · CEPT ECC (Electronic Communications Committee)

FOXCONN · GlobalLogic · Honeywell · Connecting things Hongdian · Lenovo · Linaro · MARVELL

NORCO · nuvoTon · NXP · RAD (Your Network's Edge) · Rambus · RANCHER · Rockchip 瑞芯微电子 · Red Hat

SoftBank · SolidRun (Embedded Edge Computing) · SUSE · Telco Systems (A DATA Company) · ubuntu (Delivered by Canonical) · vmware · vodafone · wipro

AMPERE · ASUS IoT · bamboo · cadence · CyberTAN · GIGABYTE · Lanner · MEDIATEK

arm SystemReady

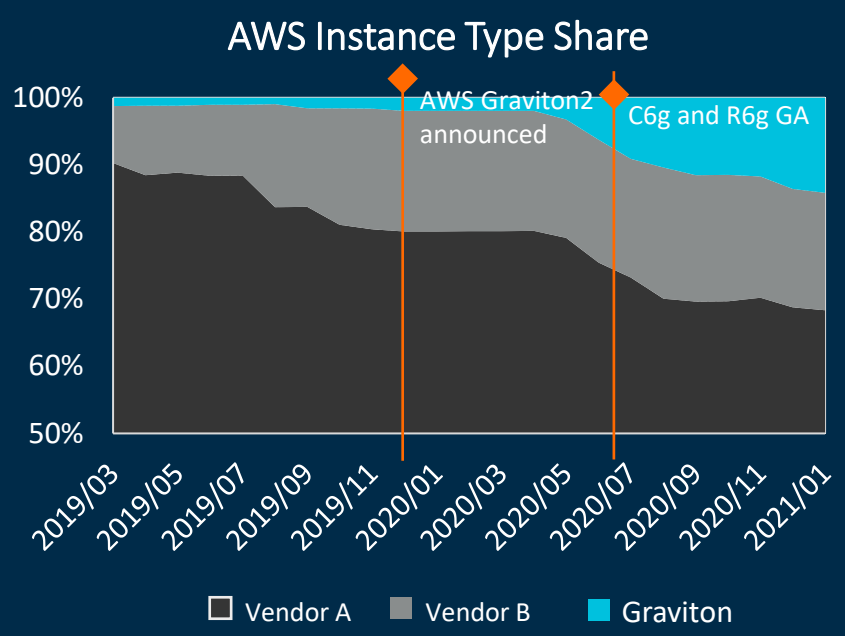
Microsoft · NEXCOM · NUVIA · OpenGCC · PEGATRON · PINE64 · Raspberry Pi · SIPEARL The Silicon Pearl

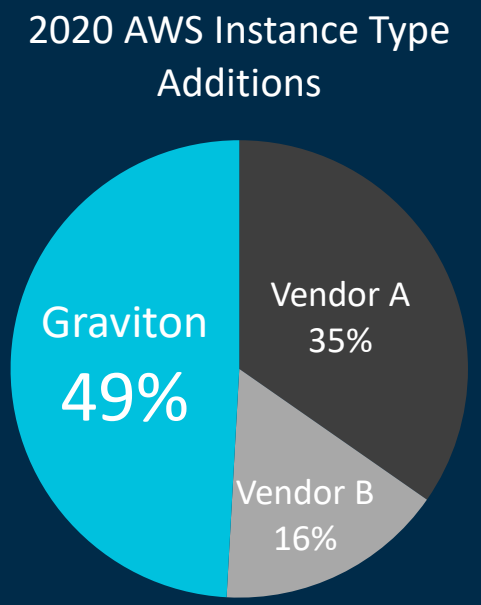
SYNOPSYS · wiwynn

Enabling a Frictionless Cloud-Native Developer Experience

AWS Graviton2 is Quickly Expanding its EC2 Footprint

Steady growth and regional expansion since mid-2020





AWS Graviton2-based EC2 instances make up 14% of the installed base within AWS

Source: Liftr Insights

49% of AWS EC2 instance additions in 2020 are based on AWS Graviton2

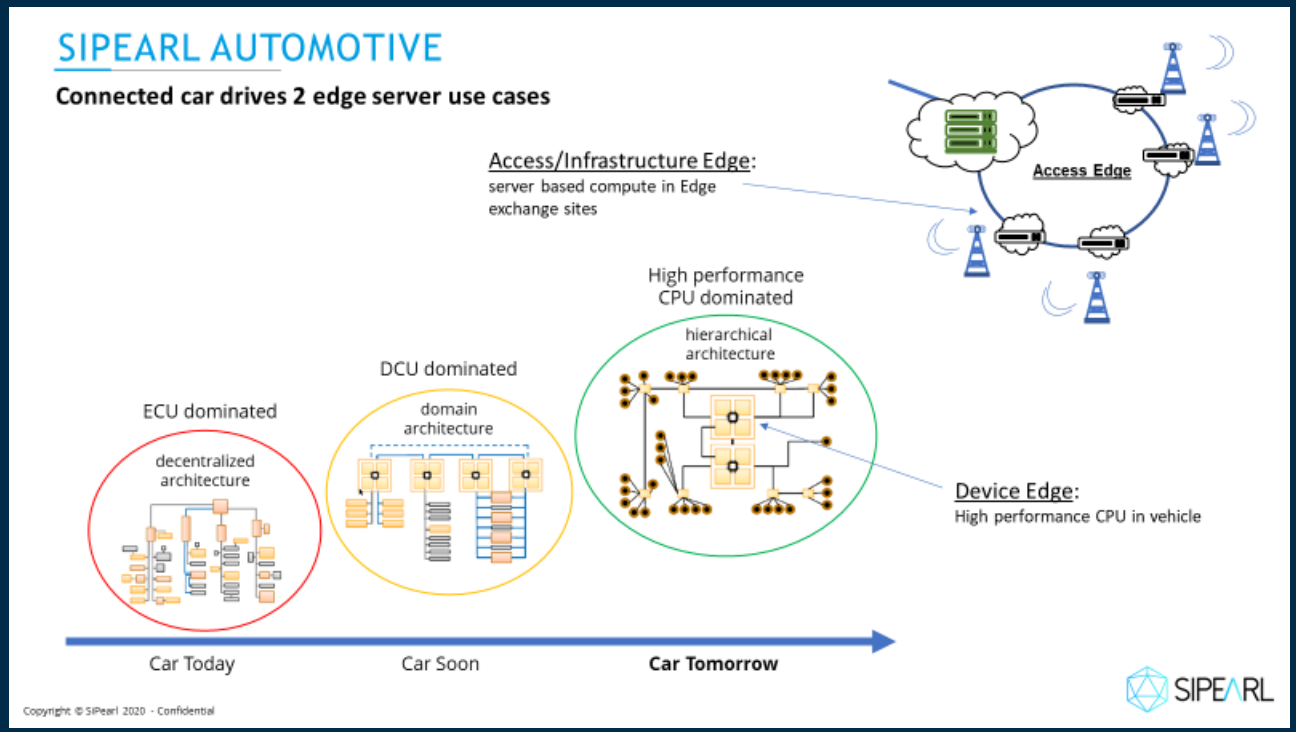
EPI: HPC IN A CAR WORK PACKAGE SIPEARL/BMW/INFINEON...

« SiPearl is working on a demonstrator on Neoverse N1 SDP platform with other EPI members (BMW, Infineon…)

- N1 SDP will be embedded in a car and will perform trajectory planning using all sensor data provided by other systems.
- Trajectory information will be sent to head unit for driver information.

We are using open source frameworks like ApolloAuto or Autoware.Ai to evaluate perfomance of those algorithm on N1 cores,

The data collected during this demo will be used to optimize those algorithm for our next generation of products using V1 cores. »



SIPEARL

SW defined cars: HPC, from the cloud to the dashboard for an amazing driver experience !

- Autonomous systems developments are driving some of the most complex HW and SW
- An industry wide effort is ongoing to enable affordable, secure, safe and power efficient systems deployments at scale
- To keep on, the car industry has to adapt the development et deployment methodologies
- Native cloud to edge development/deployments:
  - Arm Neoverse deliver the infrastructure horsepower and ISA compatibility with the edge
    - Digital twins, CI/CD, mix-criticality, SW orchestration, live data processing, connectivity…
  - Cortex & Mali automotive solution are delivering the “hpc in a car” functionalities alongside with safety and security
  - The vibrant ecosystem is aiming at enabling at scale the SW define car developments and operation

arm NEOVERSE

Arm Neoverse V1 Platform

Brian Jeff, Senior Director, Product Management

Chris Abernathy, Distinguished Engineer, CPU

Magnus Bruce, Distinguished Engineer, CPU

John Linford, Director, HPC Engineering

Arm Tech Day | April 6, 2021

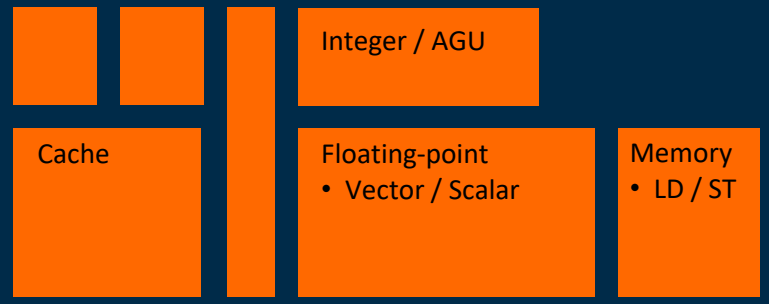
V-Series: Introducing a New Performance Tier

V-series

Neoverse V1 Platform

SVE 2x256b
bFloat16

Performance > Power & Area
WIDER / DEEPER uArch



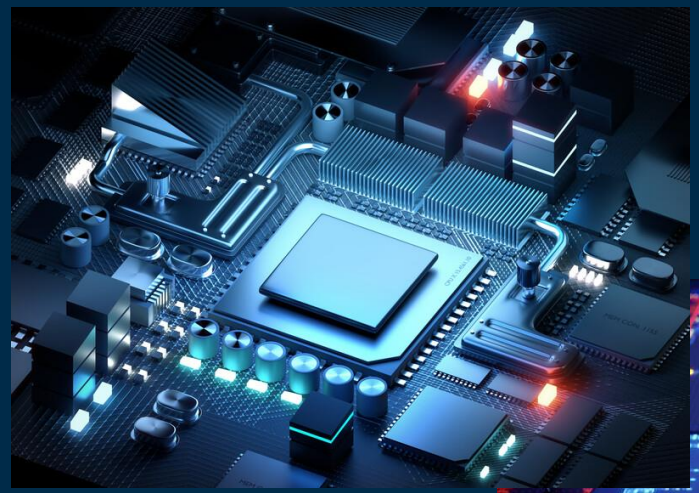
N-series

N1 Platform

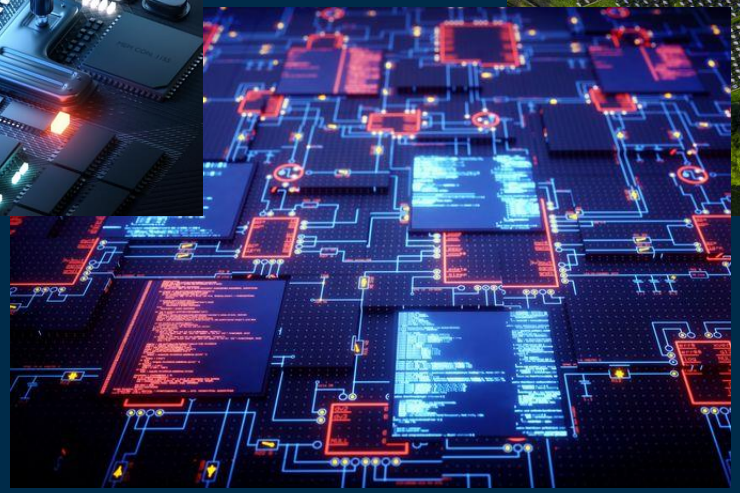
N2 Platform

Exascale System Design Objectives

CPU Per-Core Performance



Performance Efficiency



Memory Bandwidth & IO Flexibility

Technical Sovereignty

Arm Neoverse V1 Platform

A revolution in high performance computing

Arm's highest-performance core

8-wide

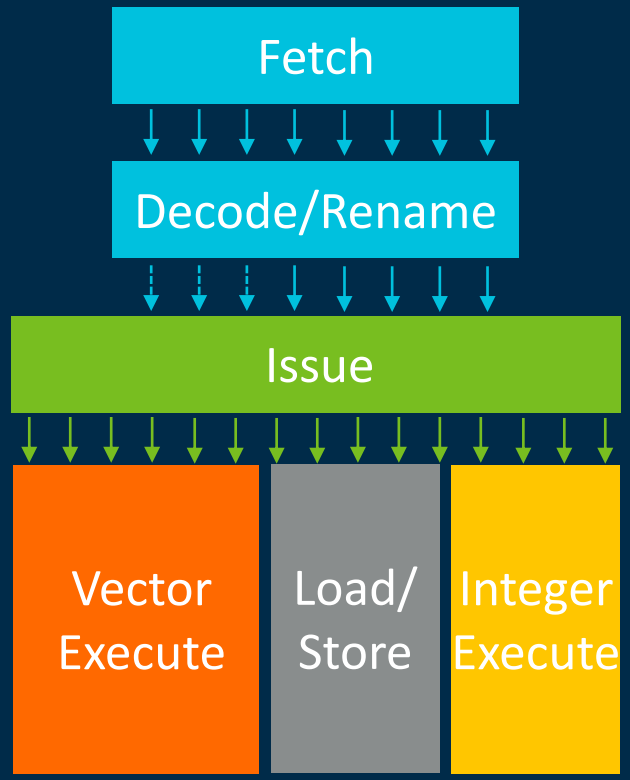
5-8-wide

15-wide

2x SVE, 4x NEON

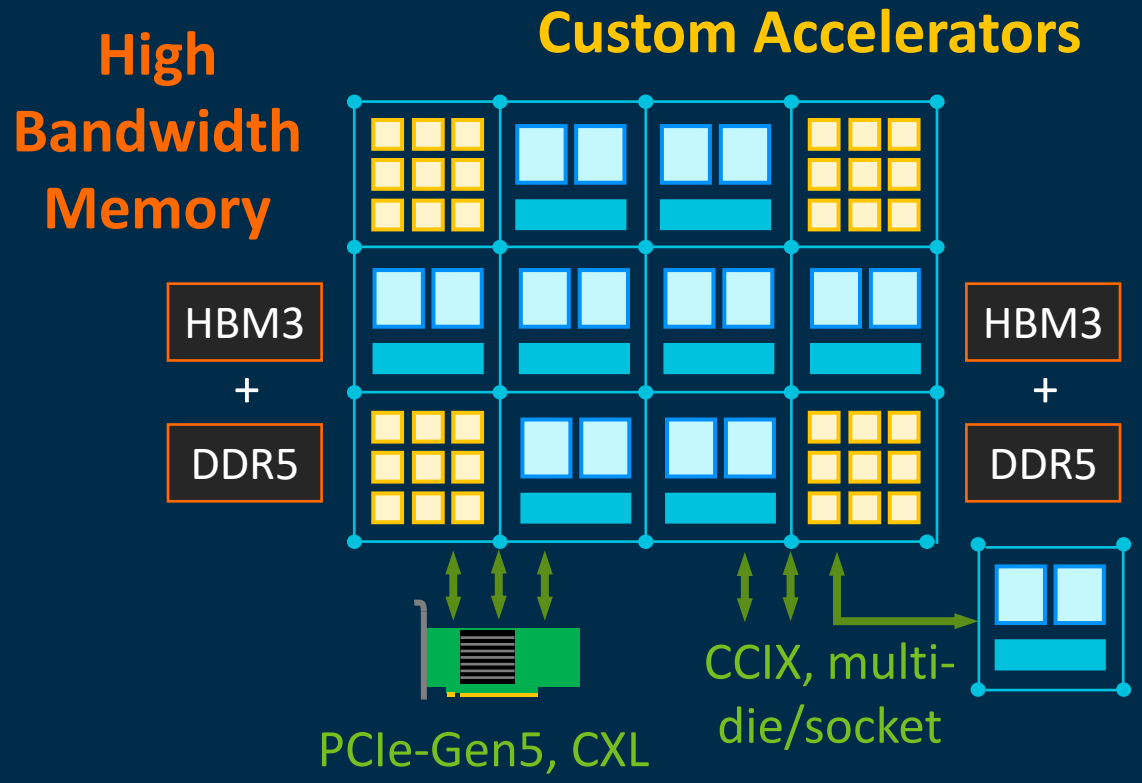
3x Ld, 2x St*

4x ALU, 2x BR



* Data; 3x Ld/St Addr

On Arm's most capable platform



Flexible IO & Multi-Chip

arm NEOVERSE

Neoverse V1 Vector Workload Gains over Neoverse N1

1.8x

Vector Workloads

2x

Floating-point performance

4x

Machine Learning performance

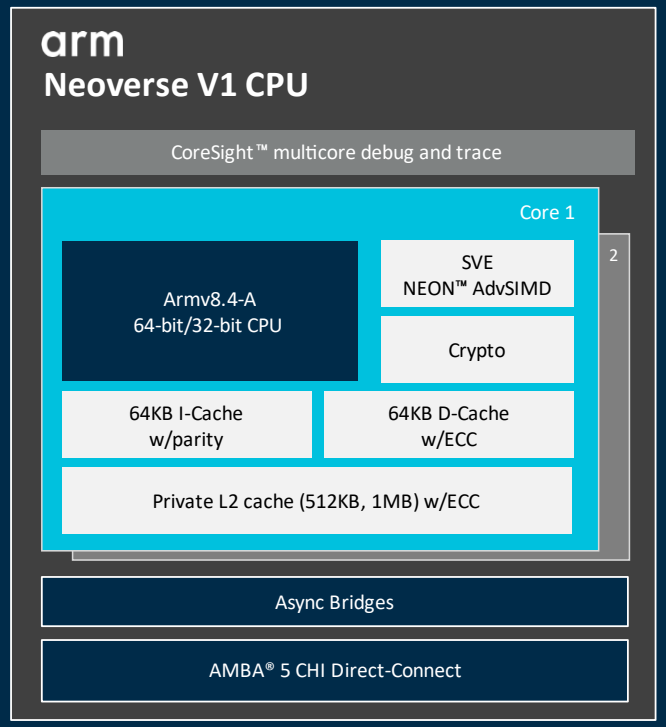
SVE

Write once, compile once, deploy forever!

What's New in Neoverse V1

Major leap above Neoverse N1

- Significant jump in performance levels
  - Micro-architectural improvements across the board
  - Improves benchmarks AND real server/HPC workloads
- Major architectural updates
  - Arm's first SVE implementation
  - Infrastructure-specific functionality, security, and performance enhancements
- Enhance scalability features
  - Performance under constrained system resources
  - Power management – hitting a power/thermal envelope

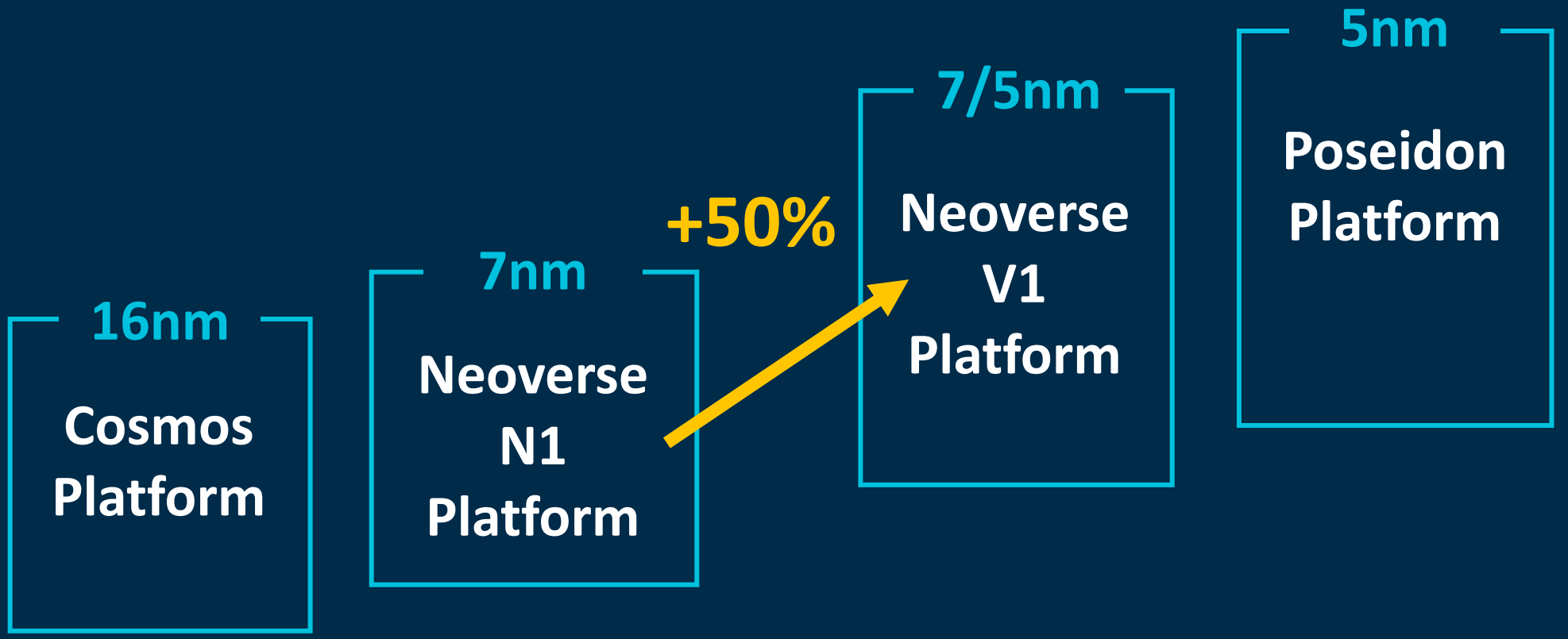


Compute Performance

Infrastructure Features

High Scalability

Continuing to Over-Deliver in Performance



30% Faster System Performance per Generation + New Features

arm NEOVERSE

Neoverse V1 Uarch Changes Since Neoverse N1

Front-End: Crucial for server and HPC workloads

- Optimized for very large instruction footprints
- Branch prediction
  - Faster ‘run-ahead’ for prefetching into the I$ (2x32B bandwidth)
  - 33% larger BTBs (8K entry)
  - 6x nano BTB (96 entry) – zero-cycle bubble
  - Generational improvements to direction accuracy
- 2x number of concurrent code regions tracked in front-end
- mprove coverage of ‘early’ branch redirects in fetch pipeline
  - Lower latency
  - Reduces costly late mispredicts

Up to **90%** reduction in branch mispredicts (for BTB misses)

Up to **50%** reduction in front-end stalls

Neoverse V1 Uarch Changes Since Neoverse N1

Mid-Core: Pushing the limits of width and depth

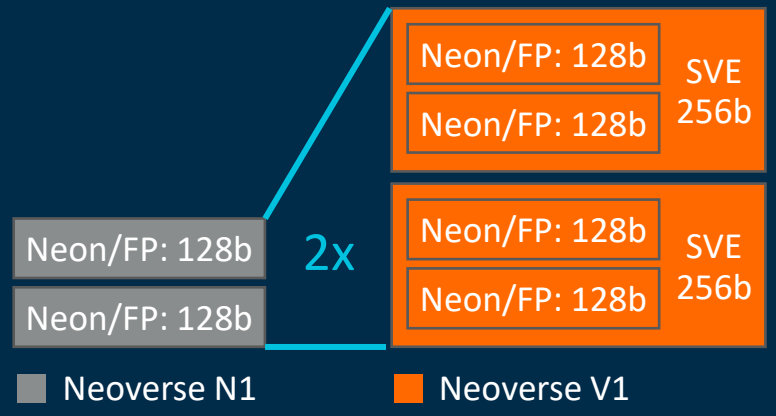
- Introduction of Mop Cache
  - L0 decoded instruction cache (3K entry)
  - High dispatch bandwidth
    - 8-instrs per cycle, 2x increase
    - I$ decode bandwidth increased from 4x to 5x per cycle
  - Lower latency decode pipeline by 1 stage
- Additional instruction fusion cases
  - Density, performance, efficiency
- OoO window size
  - 2x+ ROB (256 entry + compression)
  - Exposing parallelism (instr and mem)
- Increase superscalar integer execution bandwidth
  - 1->2 Branch Execution
  - 3->4 ALU

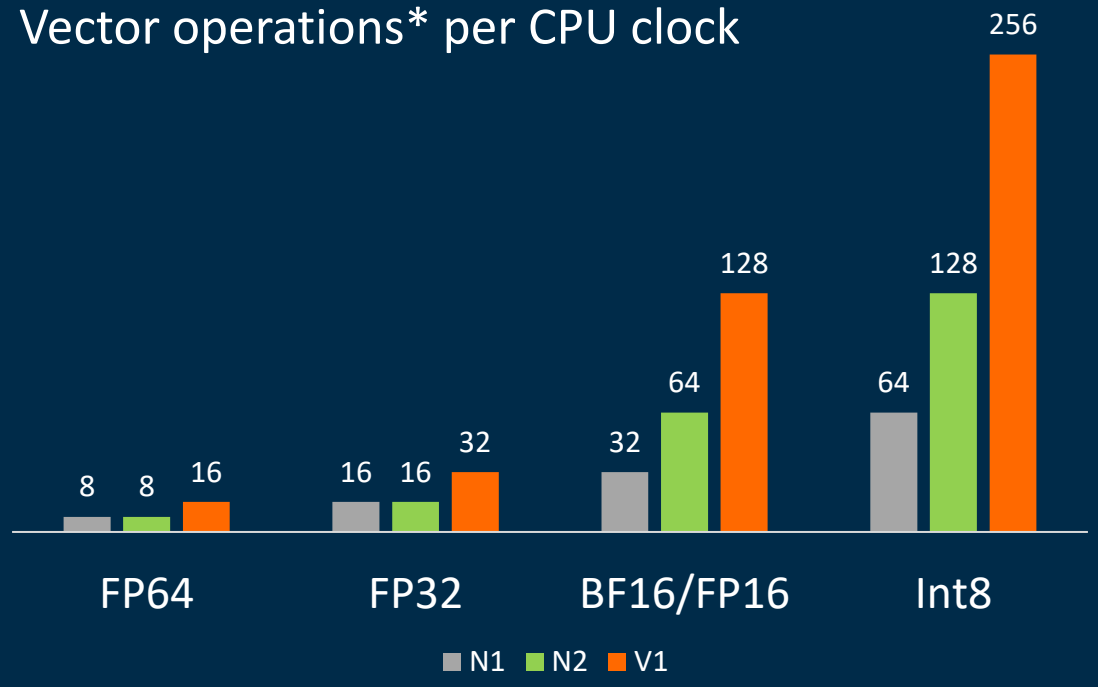
Up to **25%** increase in Integer performance for these features

Neoverse V1 Uarch Changes Since Neoverse N1

Mid-Core: Pushing the limits of width and depth

- 2x vector/fp bandwidth
  - 2x256b – SVE (new)
  - 4x128b – Neon/FP
- 4x ML performance
  - Support for bfloat16, and INT8 2x2 matrix multiply, for accelerating ML





* Multiply and accumulate is counted at two operations

arm NEOVERSE

Compute Performance

Neoverse V1 Uarch Changes Since Neoverse N1

Back-End: More bandwidth and intelligence

Bandwidth to Feed the Wide Core & SVE

- 3rd LD AGU/pipe (50% incr)
  - LS LS LD
- LD/ST data bandwidth
  - LD: 2x16B -> 3x16B
  - LD (SVE): 2x32B
  - ST: 16B -> 32B (2x), broken out into separate issue pipes

Structure Growth

- Large increase in LD/ST buffers window size
  - Exposing MLP
  - Distributed structures
- Number of outstanding external memory transactions (48->96)
  - Better latency tolerance
- MMU capacity 1.2K->2K entry (67% incr)

Up to **45%** increase in streaming bandwidth performance

arm NEOVERSE

Compute Performance

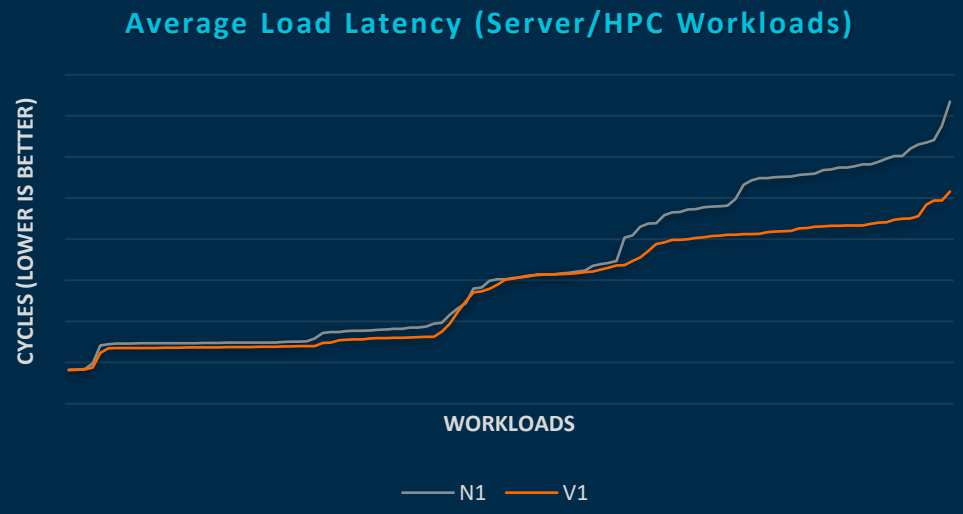
High Scalability

Neoverse V1 Uarch Changes Since Neoverse N1

Back-End: More bandwidth and intelligence

Latency Reductions

- L2 latency reduced by 1 cycle for 1M (now 10cyc load to use)
- Additional data prefetch coverage
- Improved L2 replacement policy



Scalability and Efficiency Improvements

- Dynamic prefetch behavior
  - Key to adapt aggressiveness and optimize for different systems (latency, bandwidth, congestion)
  - Better for fairness, overall throughput
- More efficient usage of SLC as victim cache

Up to **15%** reduction in L2 & SLC fills

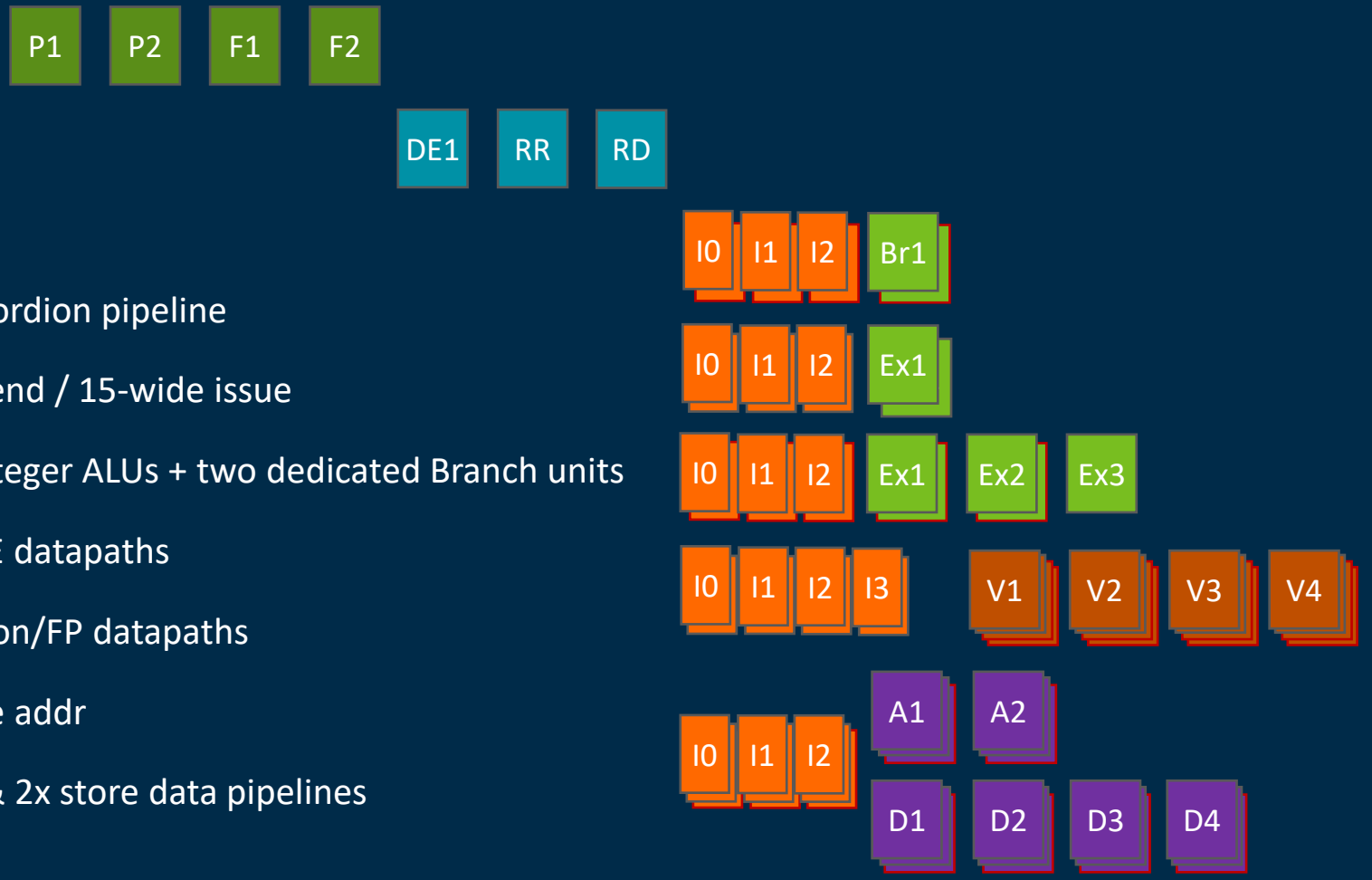
Up to **50%** reduction in L2->SLC traffic

(SpecInt2k17 Rate-N, memory-sensitive subtests)

arm NEOVERSE

Neoverse V1 CPU Pipeline

- 11+ stage accordion pipeline
- 8-wide front-end / 15-wide issue
- Four 64-bit integer ALUs + two dedicated Branch units
- 2x 256-bit SVE datapaths
- 4x 128-bit Neon/FP datapaths
- 3x load / store addr
- 3x load data & 2x store data pipelines



8-wide Instruction fetch

5-8 wide decode / rename

Branch

Integer datapath

Integer datapath
(incl. mul / div / CRC)

Vector/FP datapath

Address generation

Load / store data

Infrastructure Features

High Scalability

Neoverse V1 Power Management

Tools to better fit/fill the power/thermal envelope

Max Power Mitigation Mechanism (MPMM)

- on't have to provision systems for the worst-case CPU power
- Allows running at max frequency when in high core counts
- ntroduce 3 different 'gears' to allow for a wider range of throttling for a given workload profile

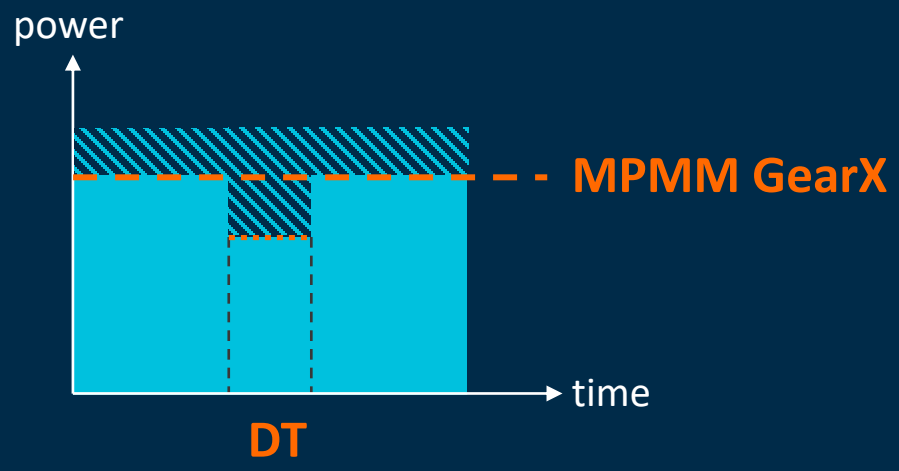
Throttling power viruses

Throttling high-bandwidth vector workloads

Aggressive throttling vector/fp workloads, with minimal integer impact

Dispatch Throttling (DT)

- For more aggressive throttling, when needed beyond MPMM
- Engaged in periods of high IPC activity (Int + Vec/FP)



arm NEOVERSE

Neoverse V1 PPA

Relative to N1 (ISO Process)

Peak performance

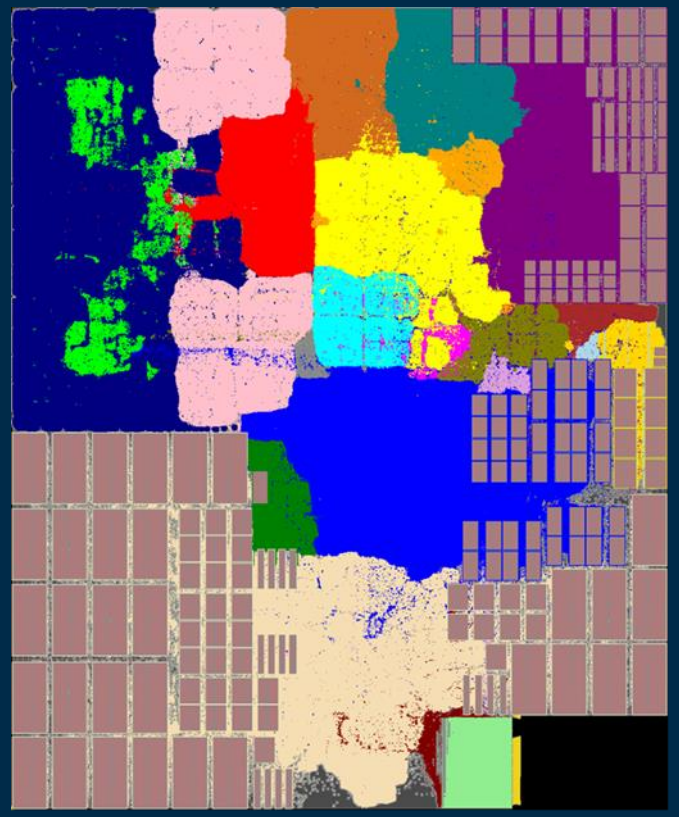
- IPC: 1.5x
- Frequency capability: 1x

Power-performance trade-off

- Power efficiency: 0.7x-1x
- *(pwr eff = perf incr / power incr)*

Performance cost

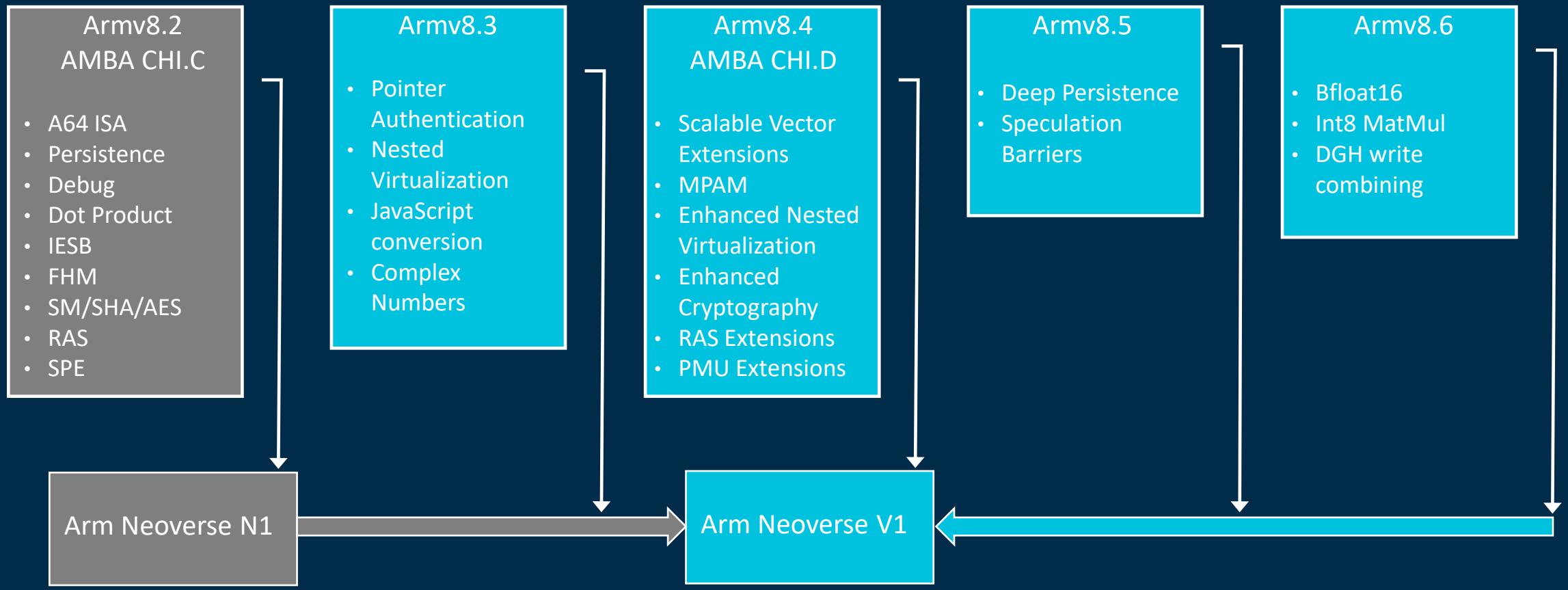
- Area – 1.7x



Example Neoverse V1 CPU die-plot – core + 1M L2

Arm Architectural Features in Neoverse V1

NEW Architecture Features



arm NEOVERSE

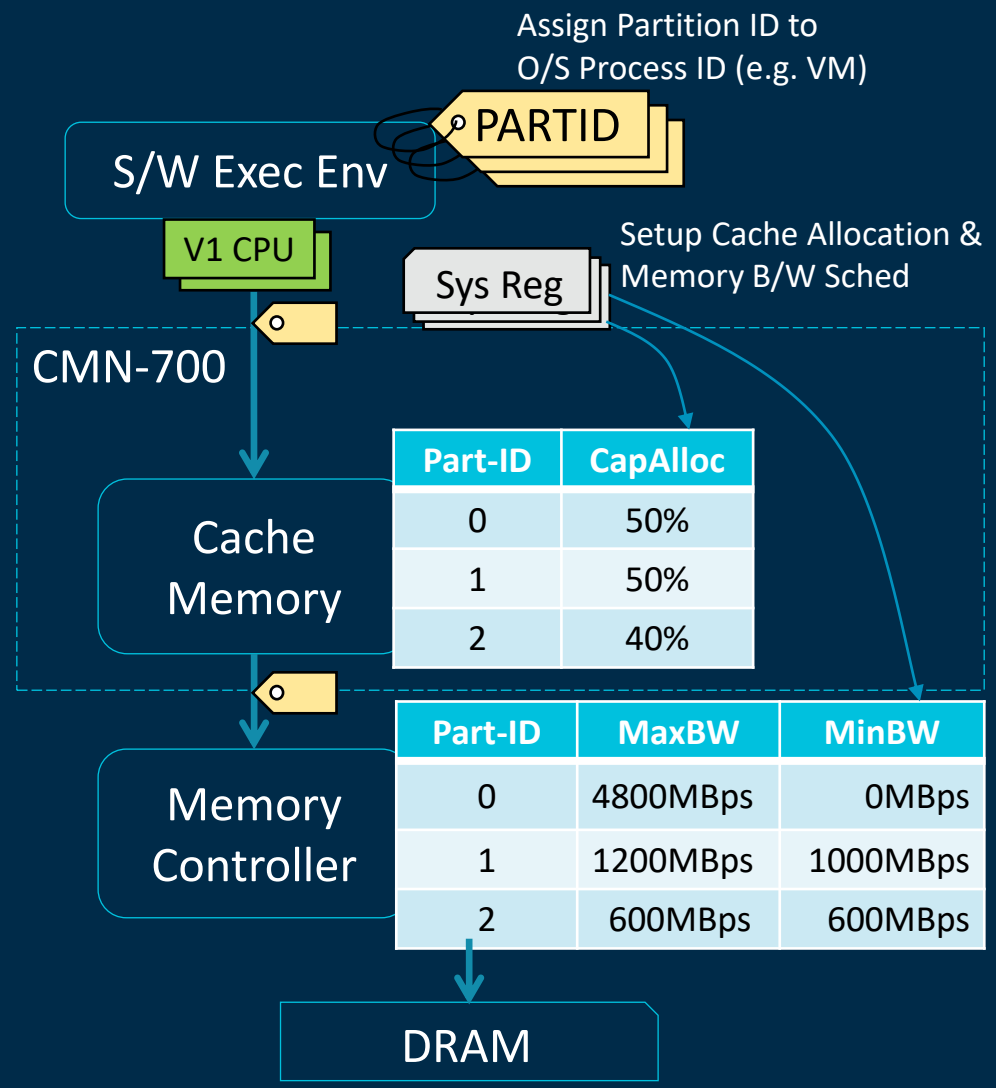
Infrastructure Features

High Scalability

Scalability Features - MPAM

Memory Partitioning and Monitoring bounds process interaction and interference in shared resources

- Control and monitoring of shared system resources
- CPU assigns Partition ID (9 bits) and Performance Monitoring Group
- Memory System Components provide controls for capacity and bandwidth
- Monitoring functions provides measurements



arm NEOVERSE

Scalability Features - CBusy

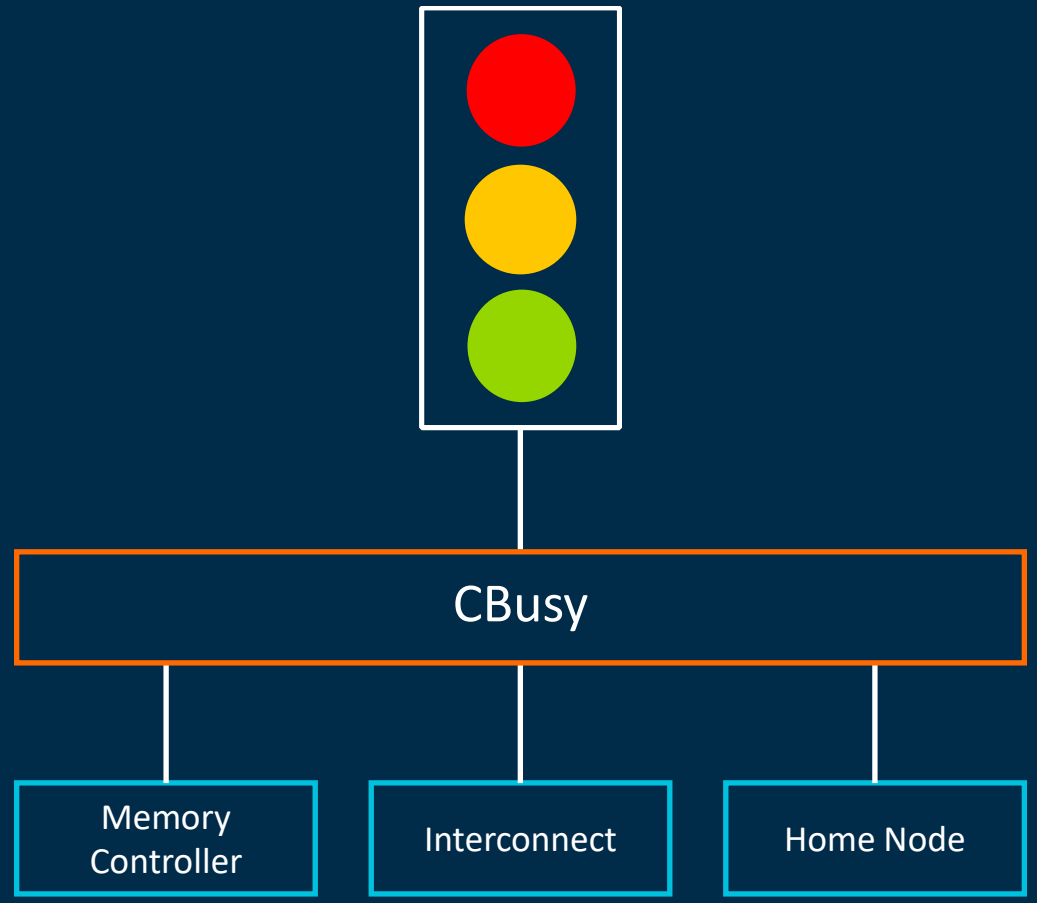
Infrastructure Features

High Scalability

Completer Busy (CBusy)

- Automatic regulation of CPU request traffic based on system congestion
- CPU aggregates and filters feedback from system
- Feedback affects hardware prefetcher aggressiveness
- CPU can also throttle its maximum outstanding transactions
- CBusy mechanisms seek a ~75% utilization of system queuing resources

15% performance uplift observed using CBusy in reference design



Performance Features

Nested Virtualization	Efficient deployment of multiple virtual environments Enables multiple Guest Hypervisors at EL1 under control of Host Hypervisor at EL2
New data types and instructions	Accelerate ML workloads Int8 and BFloat16 data types with dot product, and matrix multiply-accumulate Complex number multiplication and addition
Data Gathering Hints	Optimizes writes to I/O devices Indicates end of write gathering without the cost of a full barrier
Support for relaxed memory consistency models	New instructions to support Release Consistency processor consistent (RCpc) Enables reordering of Store-Release followed by Load-Acquire to a different address
Additional load and store atomicity	Cacheable loads and stores within an aligned 16-byte region are performed atomically
Deep persistence	Two levels of persistence for efficient management of non-volatile memory Optimized AMBA CHI-D transaction flow separates coherency and persistence

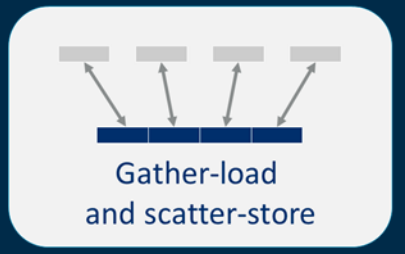
Scalable Vector Extension (SVE)

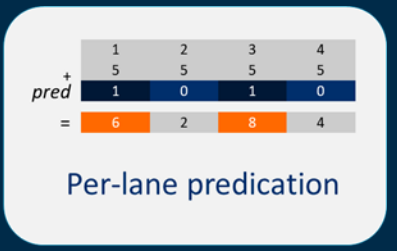
Neoverse V1 is Arm's first SVE implementation

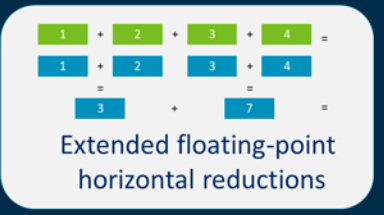
- V1 double the computation bandwidth of N1 - 2 x 256 bit
- Includes support for 4x128 Neon

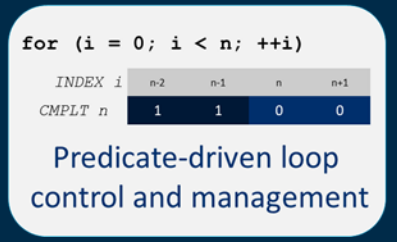
Next generation SIMD ISA

- Scalable vector length
- Fault-tolerant speculative vectorization
- Predicated execution
- Broad set of vector instructions









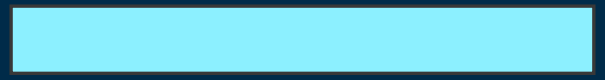
Neoverse V1 for HPC/ML

The Scalable Vector Extension (SVE)

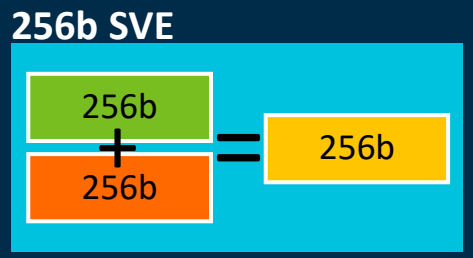
The vector engine of choice for Neoverse cores

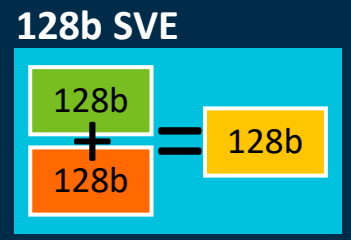
The hardware sets the vector length

In software, vectors have no length



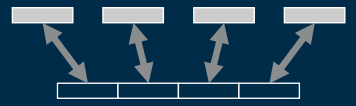
The *exact same* binary code runs on hardware with different vector lengths



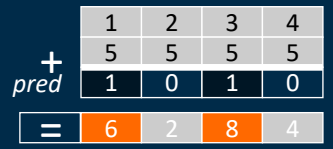


SVE: Future-Proof, Accessible, Highly Efficient

Superior auto-vectorization brings vector performance to complex codes



Gather-load and scatter-store

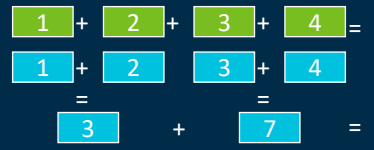


Per-lane predication

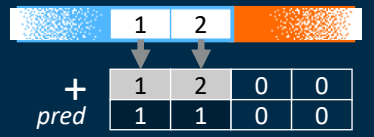
```
for (i = 0; i < n; ++i)
```

INDEX i	n-2	n-1	n	n+1
WHILELT n	1	1	0	0

Predicate-driven loop control and management



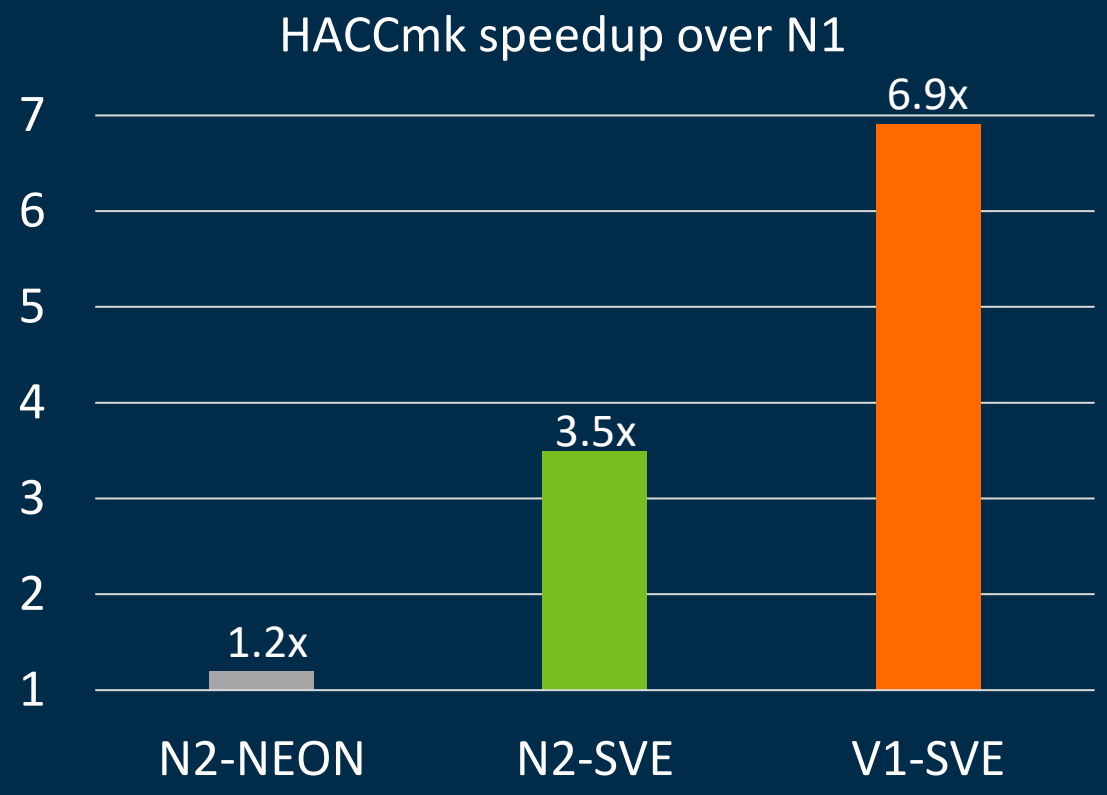
Extended floating-point horizontal reductions



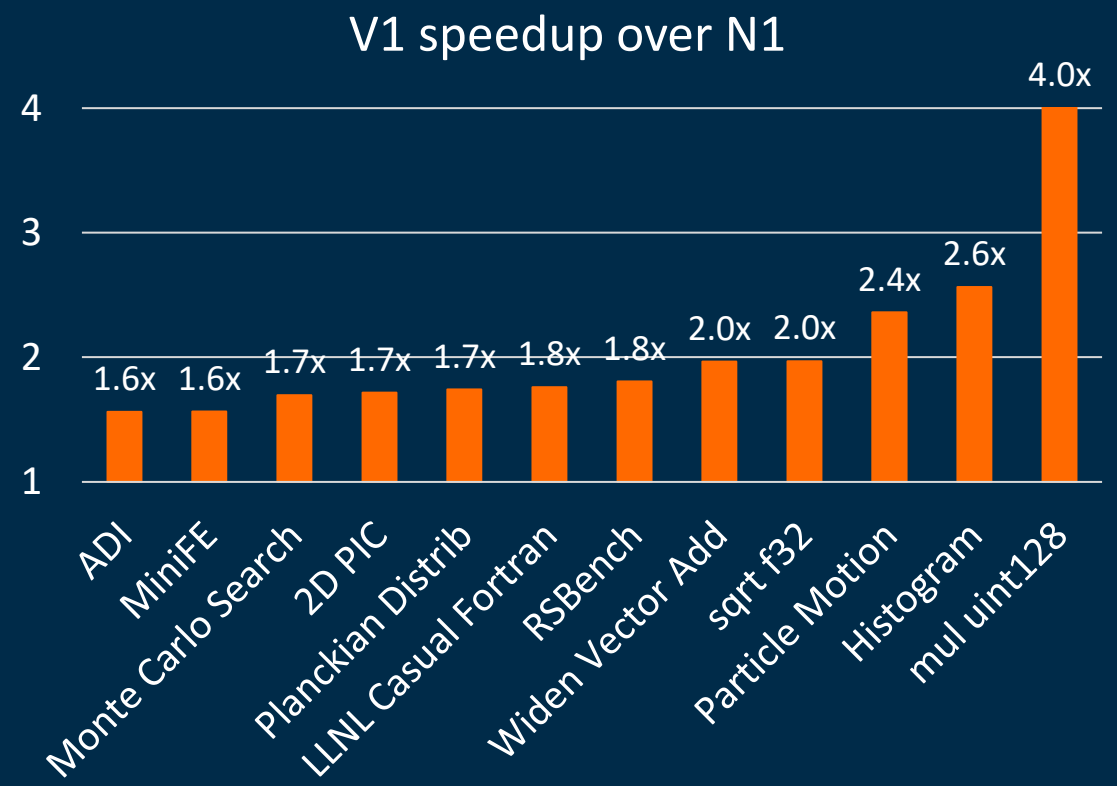
Vector partitioning and software-managed speculation

SVE in Action

Better vectorization, scalable performance



SVE improves vector code performance

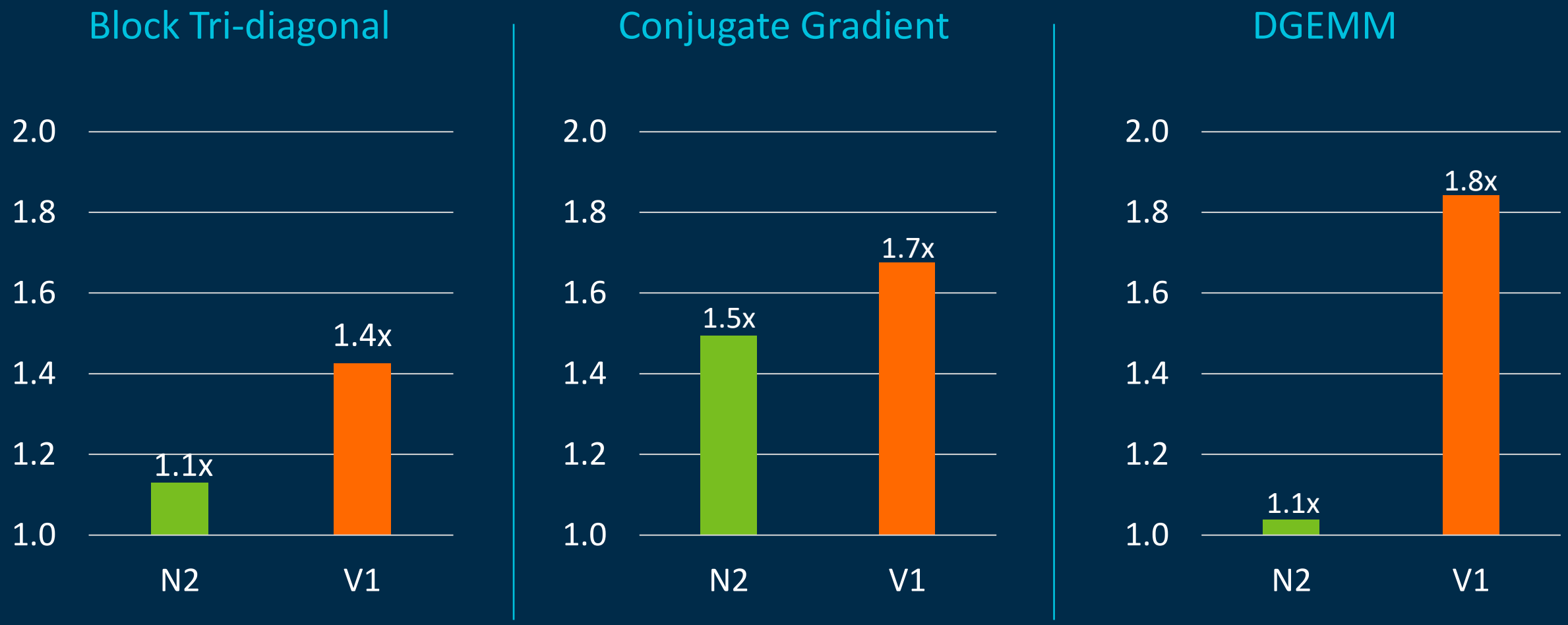


Speedups are measured in CPU cycles. Wallclock speedup is application dependent.

arm NEOVERSE

Advancing Cloud-Native HPC

NASA's NAS Parallel Benchmark Suite – Speedup over Neoverse N1

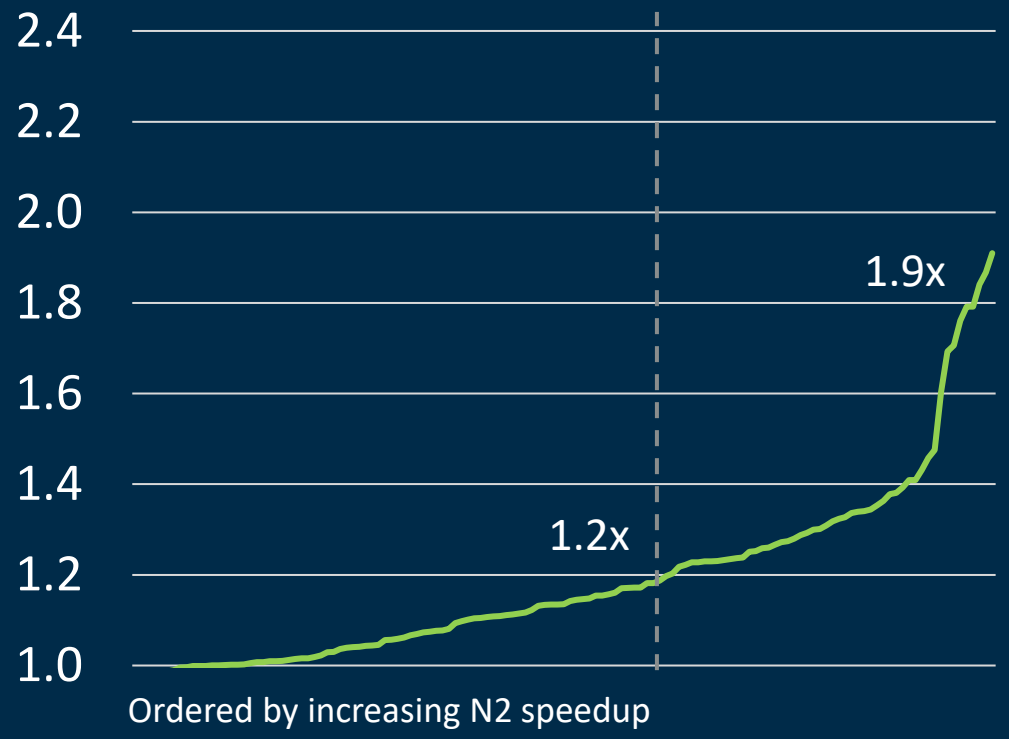


Speedups are measured in CPU cycles. Wallclock speedup is application dependent.

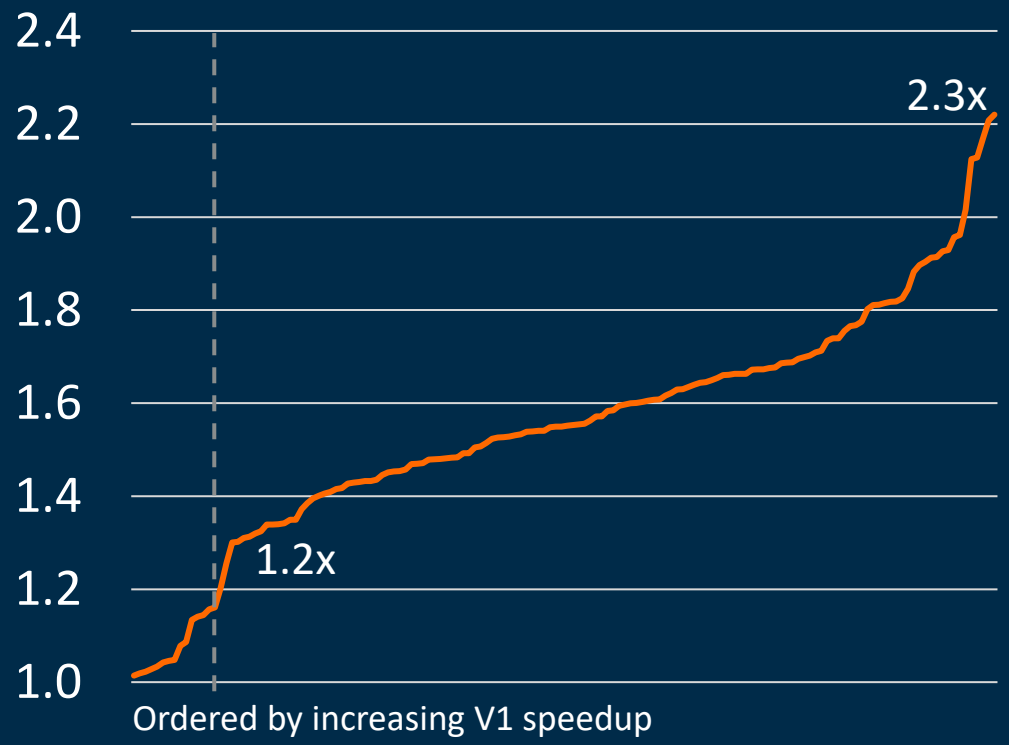
Autovectorizing Compilers for SVE

Best of GCC 10.2 or ACfL 20.3 autovectorizing TSVC-2

N2 speedup vs N1



V1 speedup vs N1



Speedups are measured in CPU cycles. Wallclock speedup is application dependent.

ACfL = Arm Compiler for Linux. TSVC = Test Suite for Vectorizing Compilers

arm NEOVERSE

Server & HPC Development Solutions from Arm

Commercially supported tools for HPC

Code Generation
for Arm servers

arm COMPILER FOR LINUX

- arm C/C++ Compiler
- arm Fortran Compiler
- arm Performance Libraries

+

Performance Engineering
cross platform, scalable

arm FORGE

arm DDT	*Debugger*
arm MAP	*Profiler*
arm PERFORMANCE REPORTS	*Reporting*

=

Server & HPC Solution
for Arm servers

arm ALLINEA STUDIO

Commercially Supported Toolkit for applications development on Linux

- C/C++ Compiler for Linux
- Fortran Compiler for Linux
- Performance Libraries
- Performance Reports
- Debugger (DDT)
- Profiler (MAP)

Neoverse V1 Platform

A revolution in high performance computing

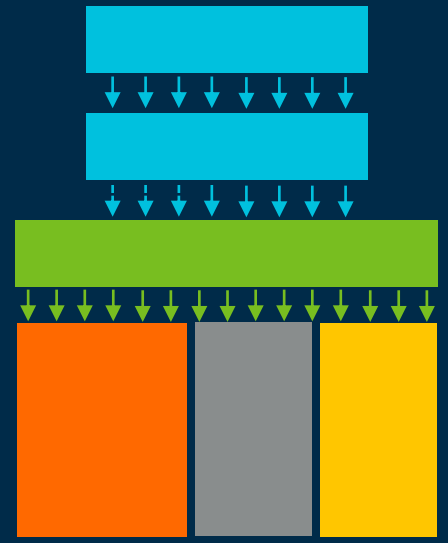
V-series

Neoverse V1 Platform

N-series

Neoverse N2 Platform

Neoverse V1 CPU



+50% IPC

2x Vector Perf

4x ML Perf

SVE

Write once, compile once, deploy forever!

HPC

HPC in the cloud

AI/ML

arm NEOVERSE

arm NEOVERSE

Thank You
Danke
Merci
谢谢
ありがとう
Gracias
Kiitos
감사합니다
धन्यवाद
شكرًا
ধন্যবাদ
תודה

arm NEOVERSE

arm NEOVERSE

Neoverse N2: Cloud-to-Edge Platforms

Mohit Taneja, Product Manager
Chris Abernathy, Distinguished Engineer
Magnus Bruce, Distinguished Engineer

Arm Tech Day | April 6, 2021

Arm Neoverse Platform PPA Design Principles

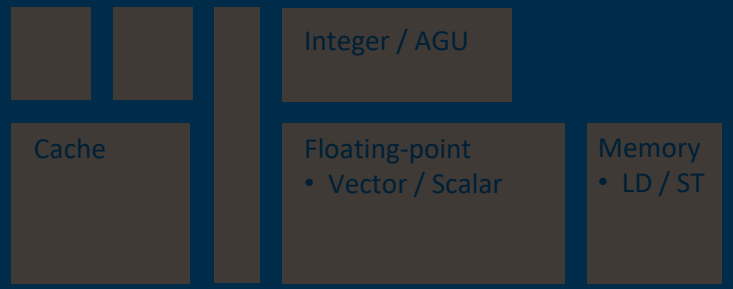
V-Series: Maximum Performance

Bigger:
- Buffers
- Caches
- Windows
- Queues

More:
- Performance
- Bandwidth
- Area
- Power

Performance > Power & Area

WIDER / DEEPER uArch



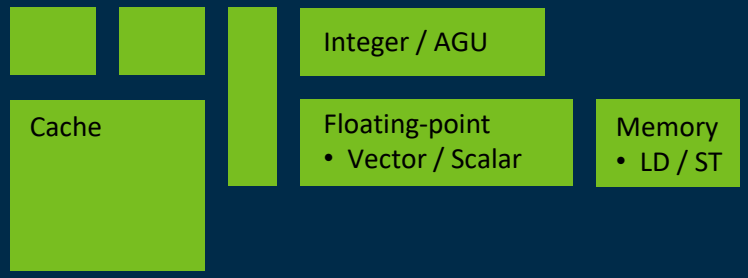
N-Series: Scale Out Performance

Optimized:
- Perf/Power
- Perf/Area

More:
- Cores per TDP/Area
- Scalability
- Balance

Performance = Power = Area

BALANCED uArch



E-Series: Efficient Throughput

Optimized:
- Power
- Area

More:
- Efficiency
- Throughput
- Thread Count

Power & Area > Performance

EFFICIENT uArch

Neoverse N2 Building on the Legacy of Neoverse N1

Massive +40% performance uplift

N-Series: Scale Out Performance

Optimized:

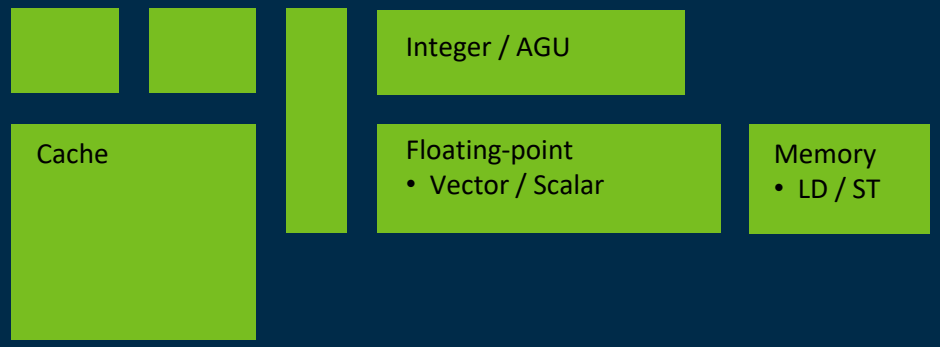
- Perf/Power
- Perf/Area

More:

- Cores per TDP/Area
- Scalability
- Balance

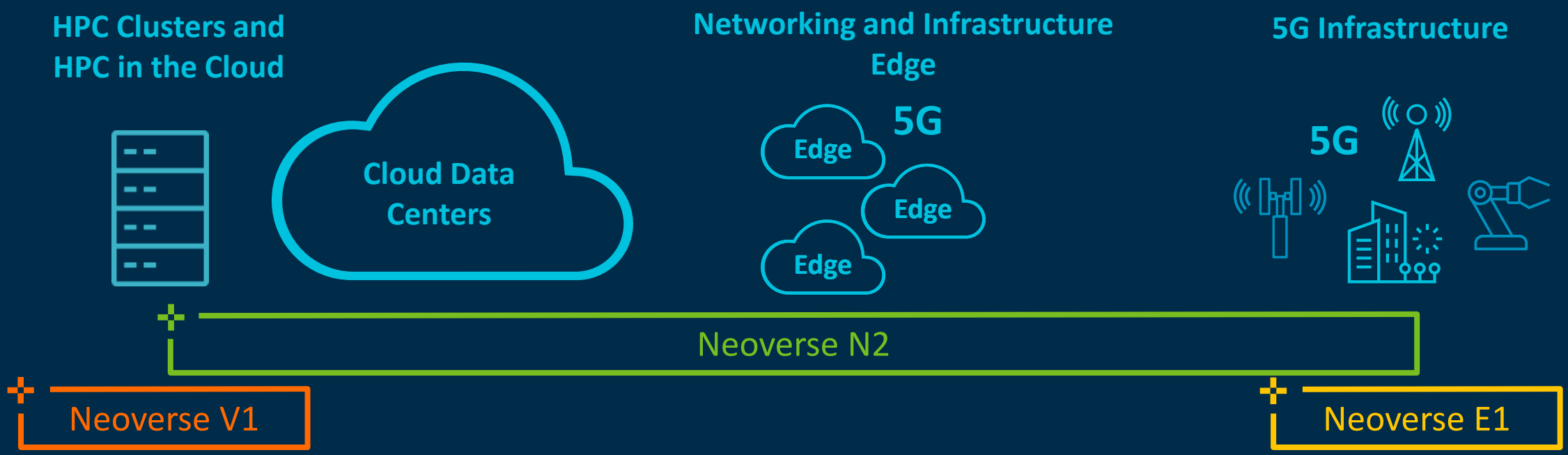
Performance = Power = Area

BALANCED uArch



- Exceptional **scalability** and performance/watt for infrastructure from the cloud to the edge
- Edge and 5G solution that delivers on **performance** and fits power & space constraints
- First Armv9 infrastructure CPU with **feature-set** designed for performance, power efficiency and security

Arm Neoverse Cloud-to-Edge Workload Positioning



32-128+ cores
80-350W TDP

High core count, high power HPC, enterprise & cloud CPUs

12-36 cores
30-80W TDP

Mid core count, mid power switch and SmartNIC CPUs

4-16 cores
20-35W TDP

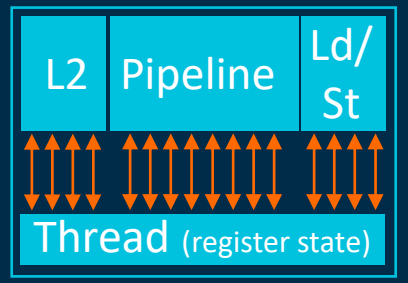
Low core count, low power gateway and router CPUs

Arm Neoverse N2 Platform

A significant uplift in cloud-to-edge performance efficiency

Cloud workloads deserve dedicated cores

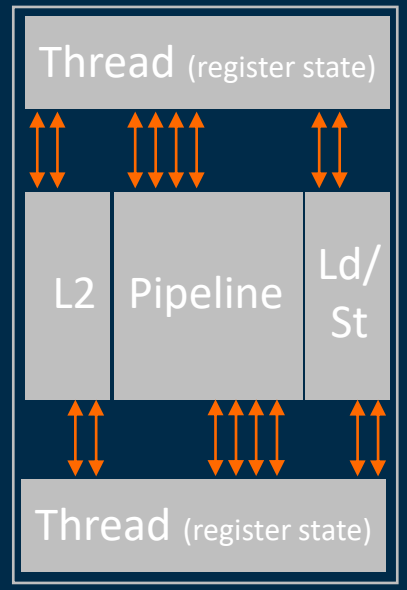
Neoverse N2



1 thread = 1 core

>>

Traditional

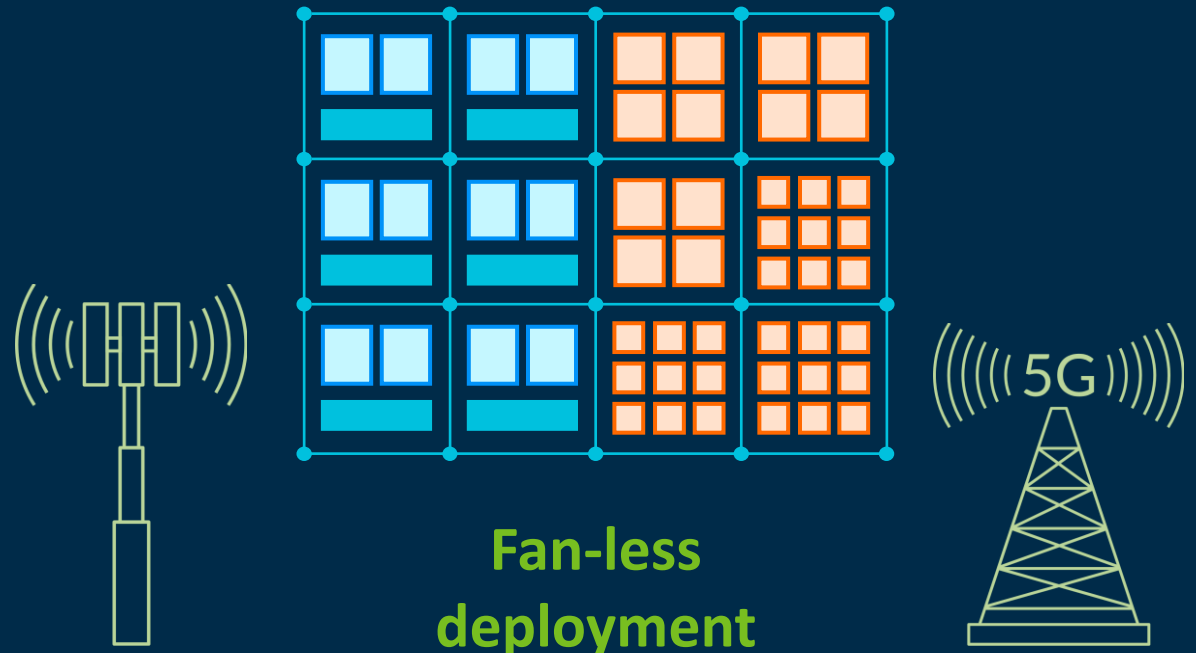


1 thread = shared core

5G needs performance and power efficiency

Performance per Watt

Heterogeneous accelerators



Fan-less deployment

Enabling Partners to Exploit Transition to 5nm

Optimal physical implementation with POP IP

Neoverse N1
on 7nm

+40%
IPC vs.
Neoverse N1

+10%
Frequency vs.
Neoverse N1

~1x
Power vs.
Neoverse N1

~1x
Area vs.
Neoverse N1

Neoverse N2
on 5nm

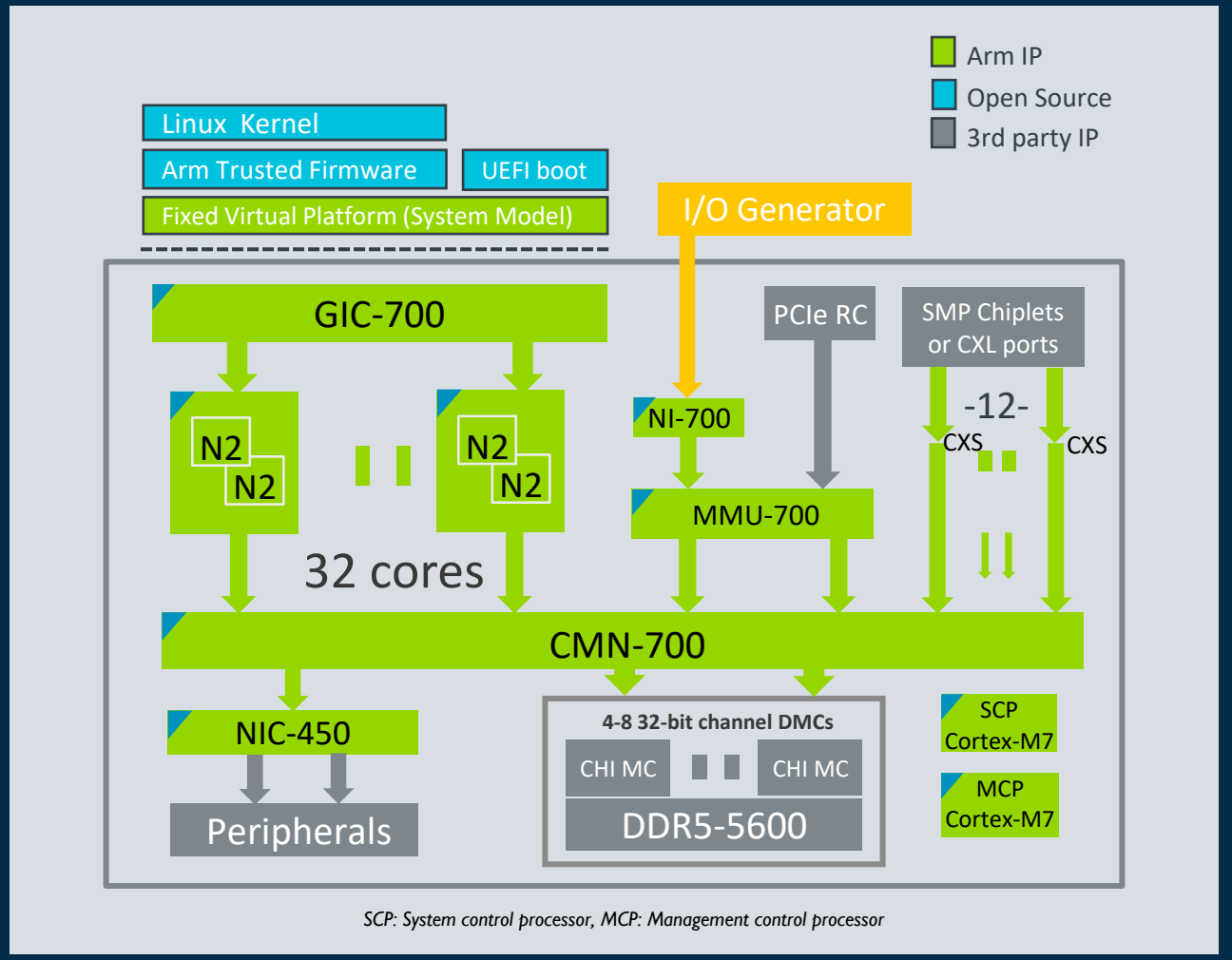
Neoverse N2 Reference Design

Targeting 5G, networking, SmartNIC and hyperscale

- Enables partners to bootstrap design
- Provides architectural guidance and best practices
- Enables left shifting software development

Included with reference design:

- PPA and benchmark results to help with sizing
- Technical guidance to build comparable system
- Fixed virtual platform for software development
- Reference software stack



SCP: System control processor, MCP: Management control processor

Neoverse N2: Market Leading Cloud-to-Edge Performance

1.4x

SPECint2006 vs Neoverse N1

1.3x

NGINX vs Neoverse N1

1.2x

DPDK L3 Fwd vs Neoverse N1

Armv9

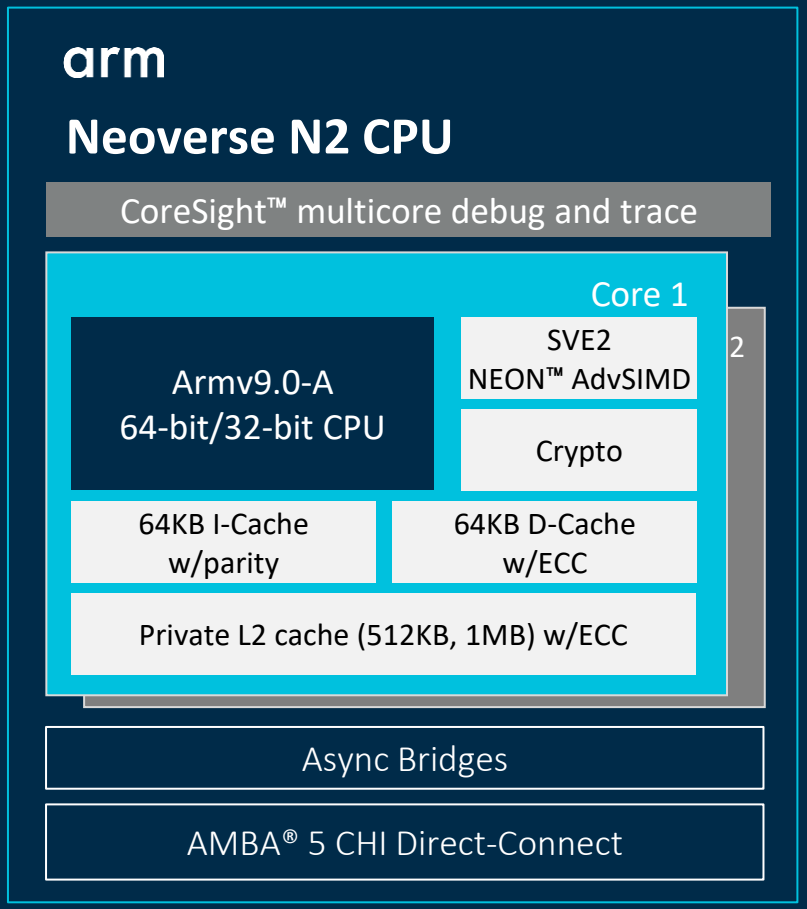
Performance, Power Efficiency and Security

*Iso-frequency performance uplifts

What's New in Neoverse N2

Major push in efficient performance over Neoverse N1

- Large jump in performance levels (+40% IPC)
  - Micro-architectural improvements across the board
  - Improves benchmarks AND real server workloads
  - Maintaining similar power- and area-efficiency as N1
  - Maximizing perf/Watt
- Major architectural updates
  - Arm's first Infrastructure v9 implementation
  - Includes latest infrastructure-specific functionality, security, and performance enhancements
- Enhanced scalability features



Efficient Performance

Extreme Scalability

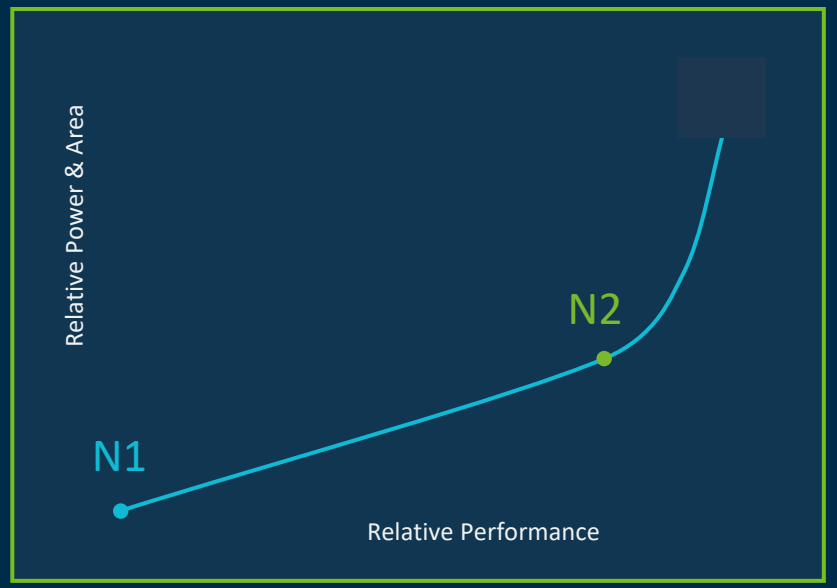
Infrastructure Features

Security

Neoverse N2 Efficient Performance

Our approach beyond N1

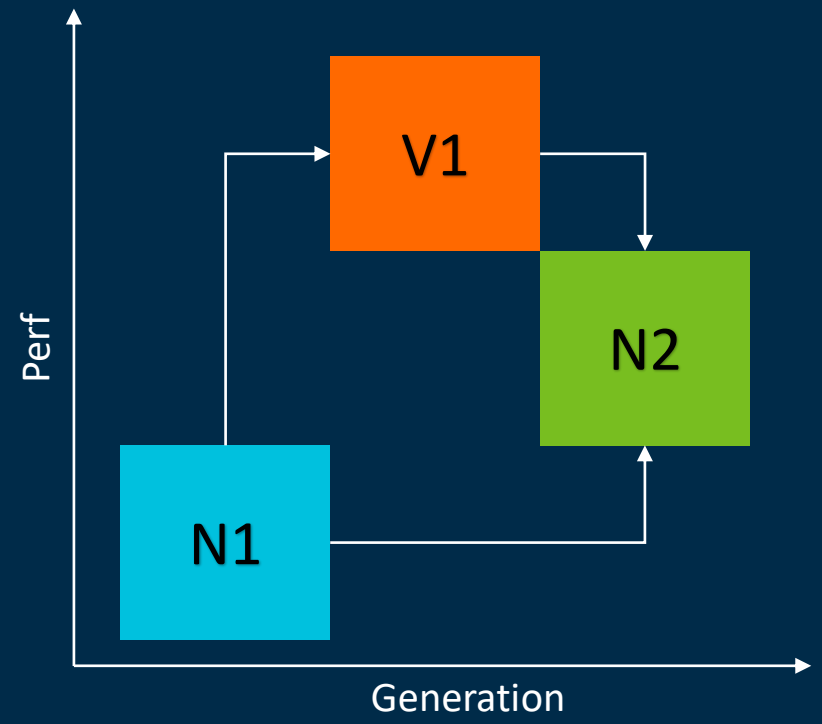
- Performance features must pay for themselves
  - Power- and area-efficiency
- Optimize efficiency of existing structures across the board
- Relative to Neoverse V1
  - Relies less on the ‘width’ and ‘depth’ performance levers in the mid-core
    - 5-wide dispatch
    - 160+ instr ROB
  - Reduced amount of speculation throughout
  - 2x128 SVE/Neon/FP datapath
  - Reduced load-store bandwidth



Neoverse N2 Efficient Performance

Similarities with Neoverse V1

- N2 inherits many of the efficient “smarts” that went into Neoverse V1
  - Branch-prediction algorithms
  - Data-prefetching algorithms
  - Replacement policies
- Front-end very similar to Neoverse V1
  - Important for infrastructure applications with large instruction footprints and complex branch behavior
  - Also has a Mop Cache for accelerating smaller kernels (1.5K entry)



Infrastructure Features

Extreme Scalability

Neoverse N2 Power Management

Tools to better fit the power & thermal envelope

- Similar capabilities from Neoverse V1, plus…

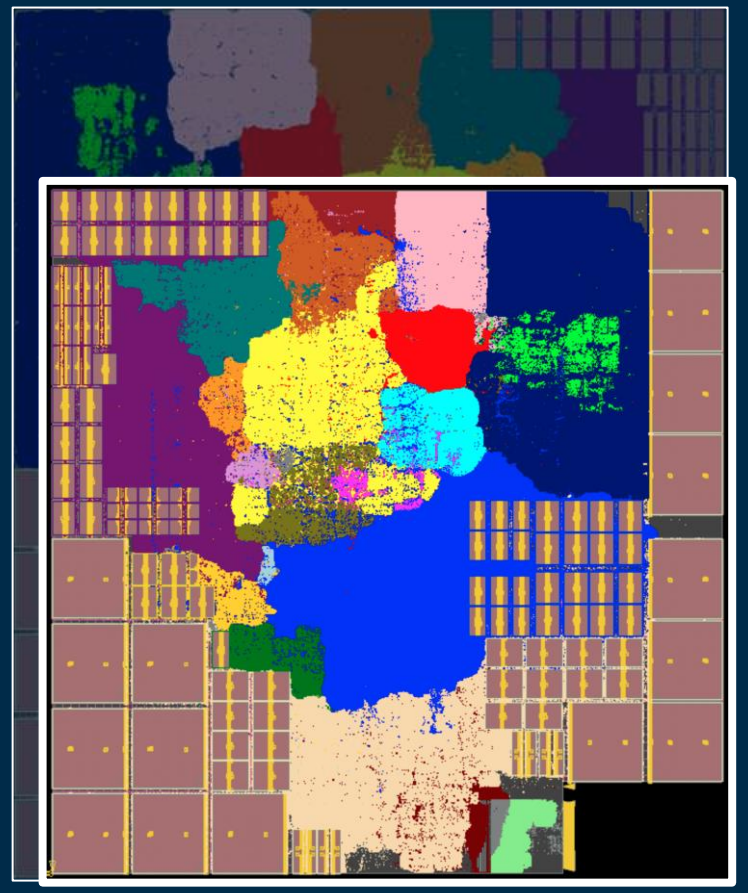
- Performance Defined Power Management (PDP)
  - Dynamic scaling of the Uarch to maximize efficiency for the given workload
    - The CPU knows how to do this best
  - Targeting width, depth, speculation (internally and externally) for scaling
    - Not every workload requires maximum of these for unleashing performance
  - High efficiency scaling compared to MPMM & DT

Neoverse N2 PPA

- Relative to N1 (ISO process & configuration)
  - IPC – 1.4x
  - Frequency capability – 1x
  - Area – 1.3x
  - Power – 1.45x

V1

N2

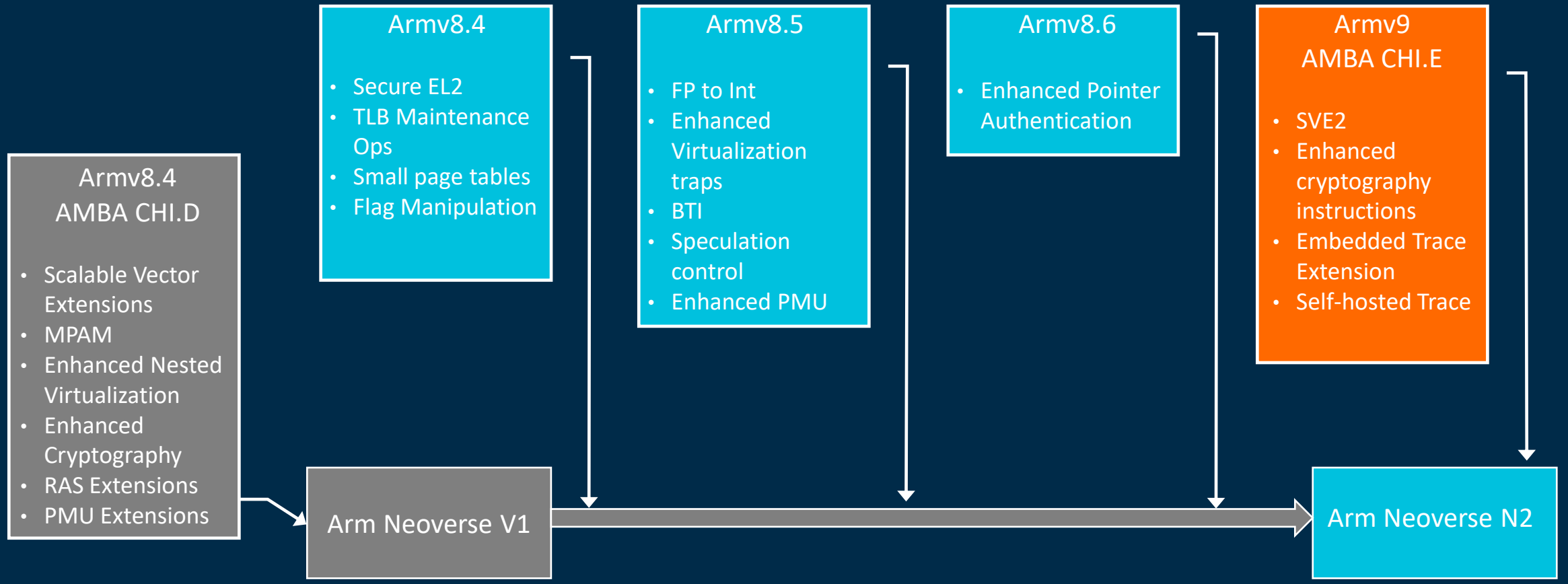


25% smaller,
more threads

Example Neoverse N2 CPU die-plot – core + 1M L2

Arm Architectural Features in Neoverse N2

NEW Architecture Features



arm NEOVERSE

Scalability Features

Extreme Scalability

Security

TLB Invalidate enhancements

- Reduced Distributed Virtual Memory traffic throughout system
- Includes TLBI domain, address range, and TLB level hints

Secure EL2

- Hypervisor management of secure resources without exposing EL3
- Provides isolation of:
  - EL3 software from Secure EL1 software
  - Normal world software from Secure EL1 software
  - Distinct Secure EL1 software components from each other

Security Features

Pointer Authentication

- Protects from Return Orientated Programming code reuse attacks
- Pointers are protected and authenticated using QARMA algorithm
- Authentication failure triggers a fault

Branch Target Identification

- Protects from Jump Orientated Programming code reuse attacks
- BTI instructions provide landing pads for indirect branches (BR, BLR, RET)
- Indirect branch to non-landing-pad location in marked pages causes exception

Memory Tagging

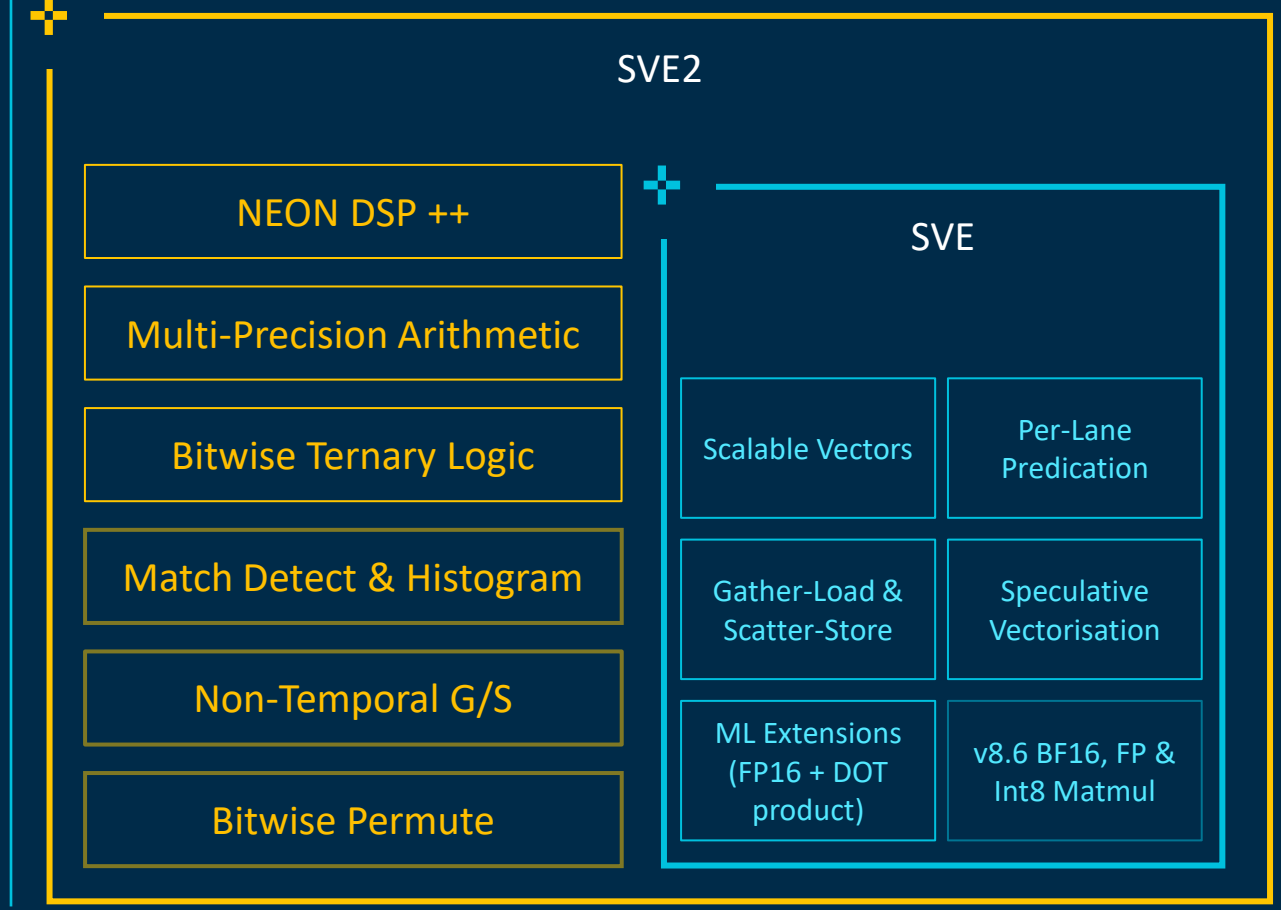
- Detects memory safety errors (e.g use after free and buffer overflow)
- Memory tagged with 4 tag bits per 16 data byte
- Pointers colored with tag value in unused upper bits
- Loads and stores check pointer tag against memory tag

Performance Features

Scalable Vector Extensions 2

- Neon + SVE
- Acceleration of more workloads
  - ML
  - DSP
  - Media
  - 5G
- Brings the benefits of auto-vectorization to new applications
- Write once, compile once, deploy forever

SVE Has Evolved



Neoverse N2: Market Leading Cloud-to-Edge Performance

Scales from Cloud-to- Edge

>128

Threads per socket

1.4x

SPECint2006 vs Neoverse N1

Armv9

Performance, Power Efficiency and Security

arm NEOVERSE


arm NEOVERSE
The Cloud to Edge Infrastructure Foundation
for a World of 1T Intelligent Devices
Thank You!

arm

Thank You
Danke
Gracias
谢谢
ありがとう
Asante
Merci
감사합니다
धन्यवाद
Kiitos
شكرًا
ধন্যবাদ
תודה

22-23-24 JUNE | DIGITAL EVENT

Forum Teratec 2021

Unlock the future !

Merci pour votre attention.

Thank you for your attention.

PLATINUM SPONSORS	Atos	ddn	GRAPHCORE	Hewlett Packard Enterprise	intel.	VAST
GOLD SPONSORS	ATEMPO	cea	doitnow HPC Services	exaion EDF GROUP	Lenovo	UCIT
SILVER SPONSORS	arm	aws	GENCI	NVIDIA.	ORACLE	rescale	XILINX.
PARTENAIRE EUROPA VILLAGE	Inria					