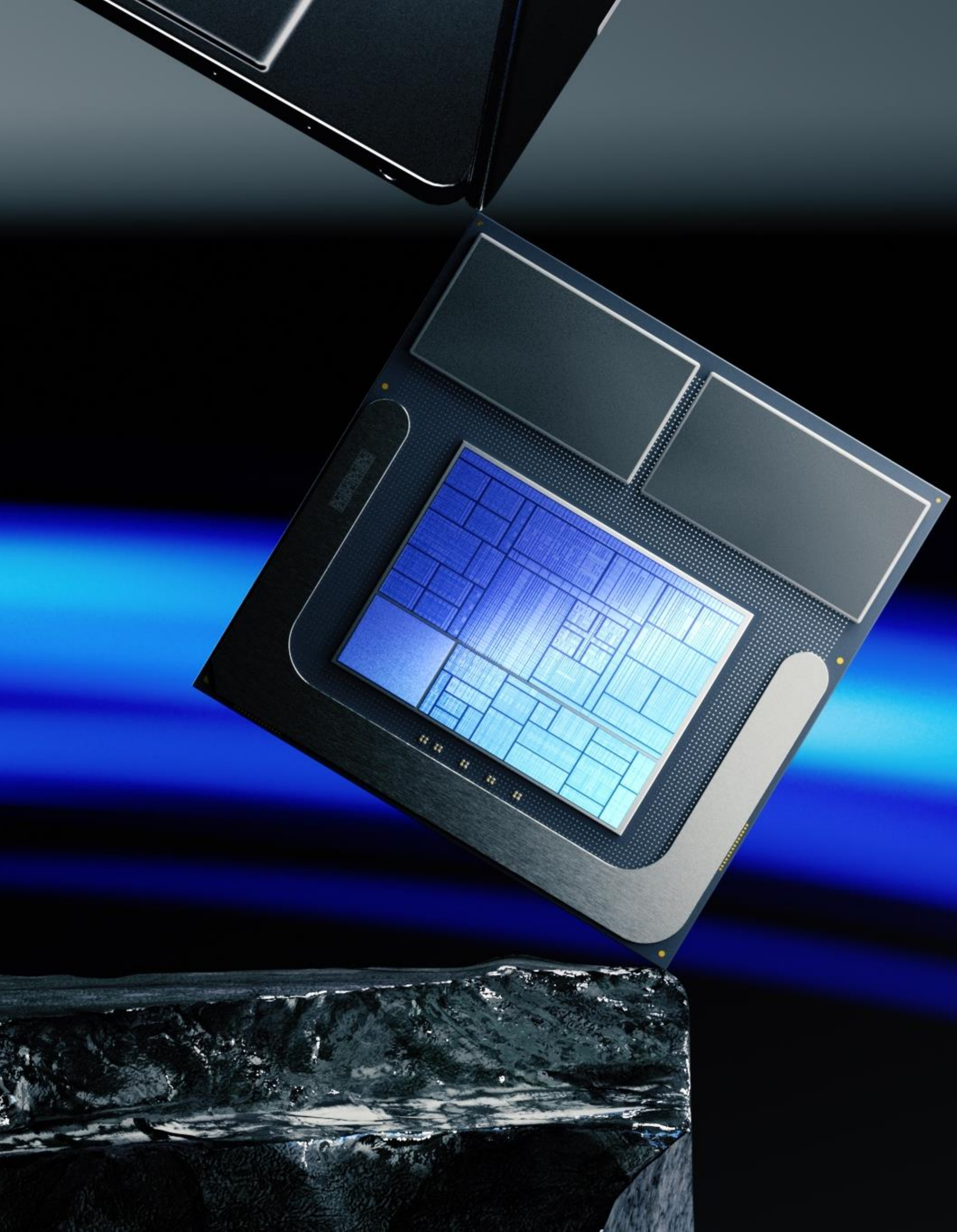


Lunar Lake Architecture Session

HotChips 2024

Arik Gihon

Senior Principal Engineer, CCG
SoC Architecture

Lunar Lake

Flagship SoC for the next gen of AI PCs

Breakthrough x86 power efficiency

Up to 40% lower SoC power*

Exceptional core performance

Similar ST perf at half the power*

Massive leap in graphics

Up to 1.5X better graphics*

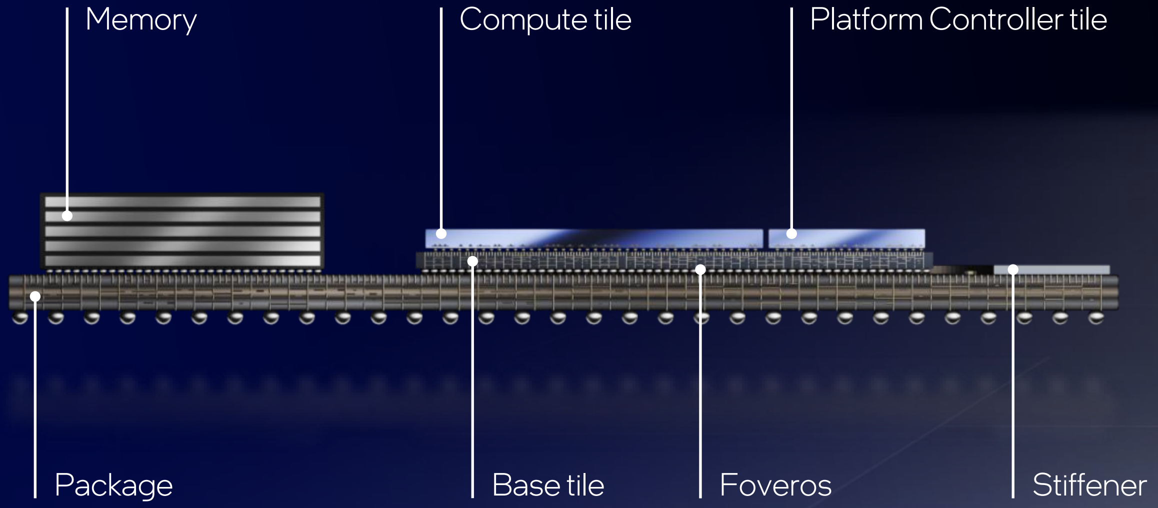
Unmatched AI compute

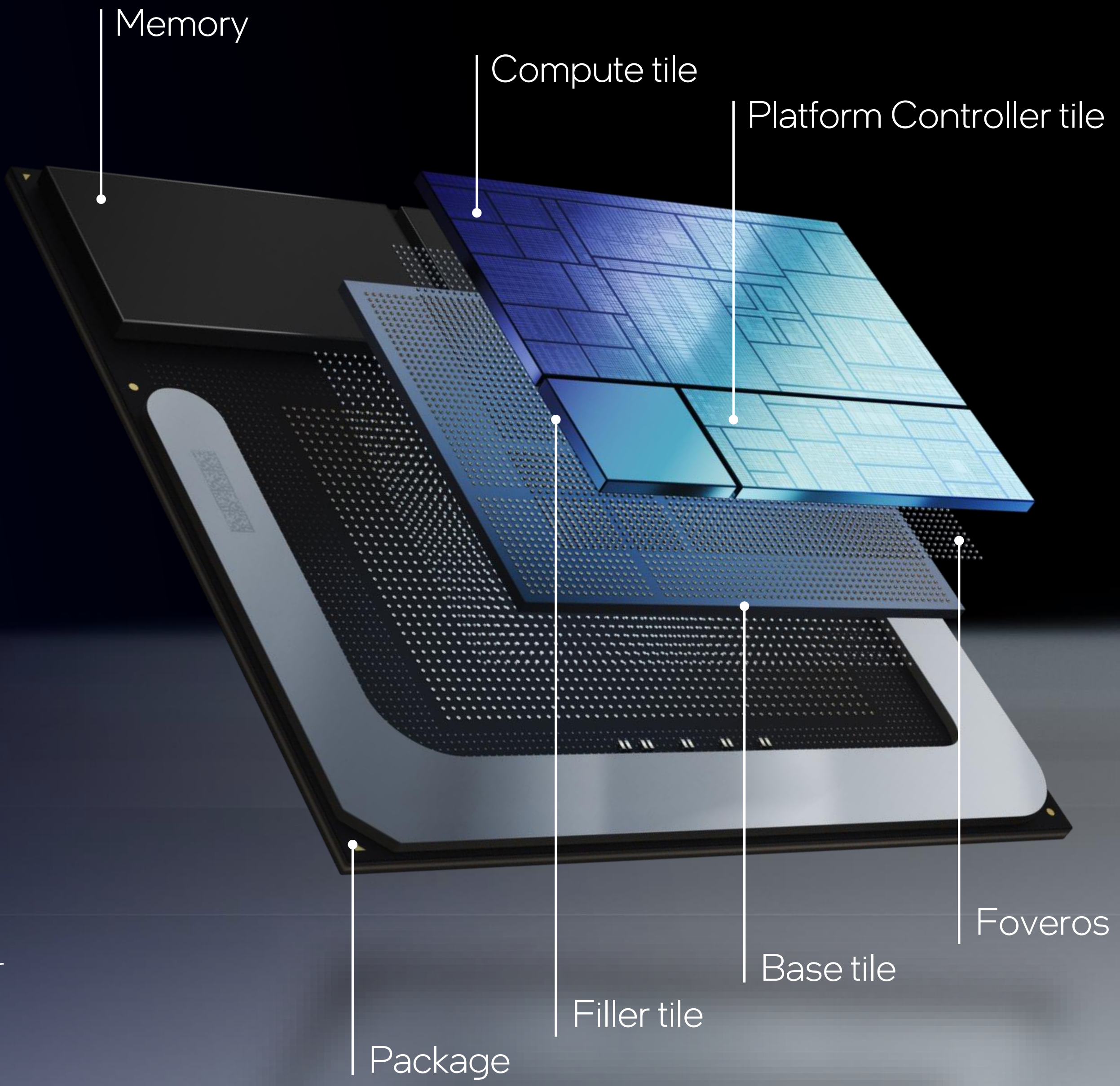
Up to 120 platform TOPS

*Versus Intel's previous gen. For more information, go to Intel.com/PerformanceIndex.

Lunar Lake Construction

Built with advanced packaging





Memory on Package

First ever Intel integration onto package

Up to
32GB
with 2 ranks

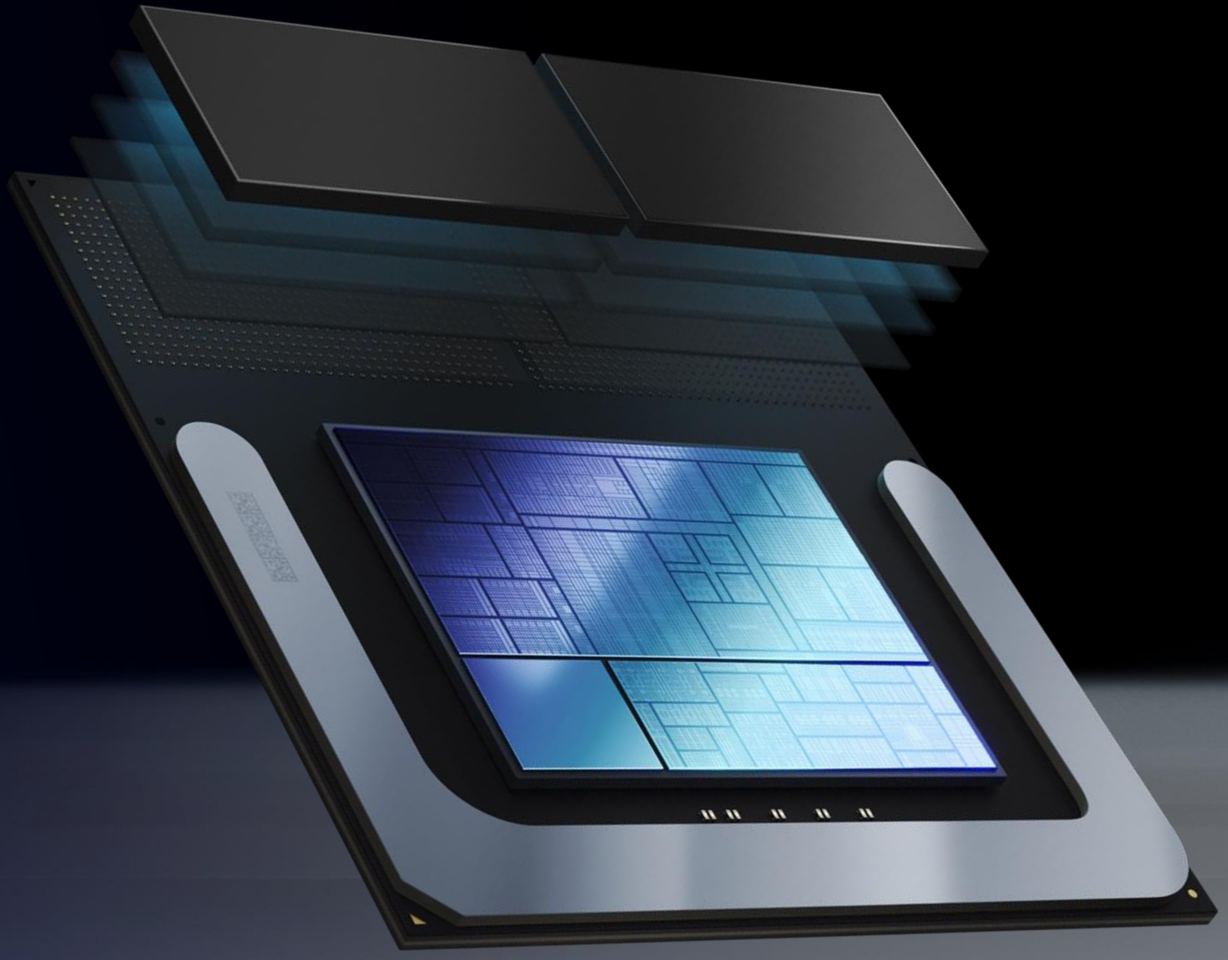
Support for
LPDDR5x
DRAM

Up to
8.5GT/s
per chip

Support for
16b x4
channels

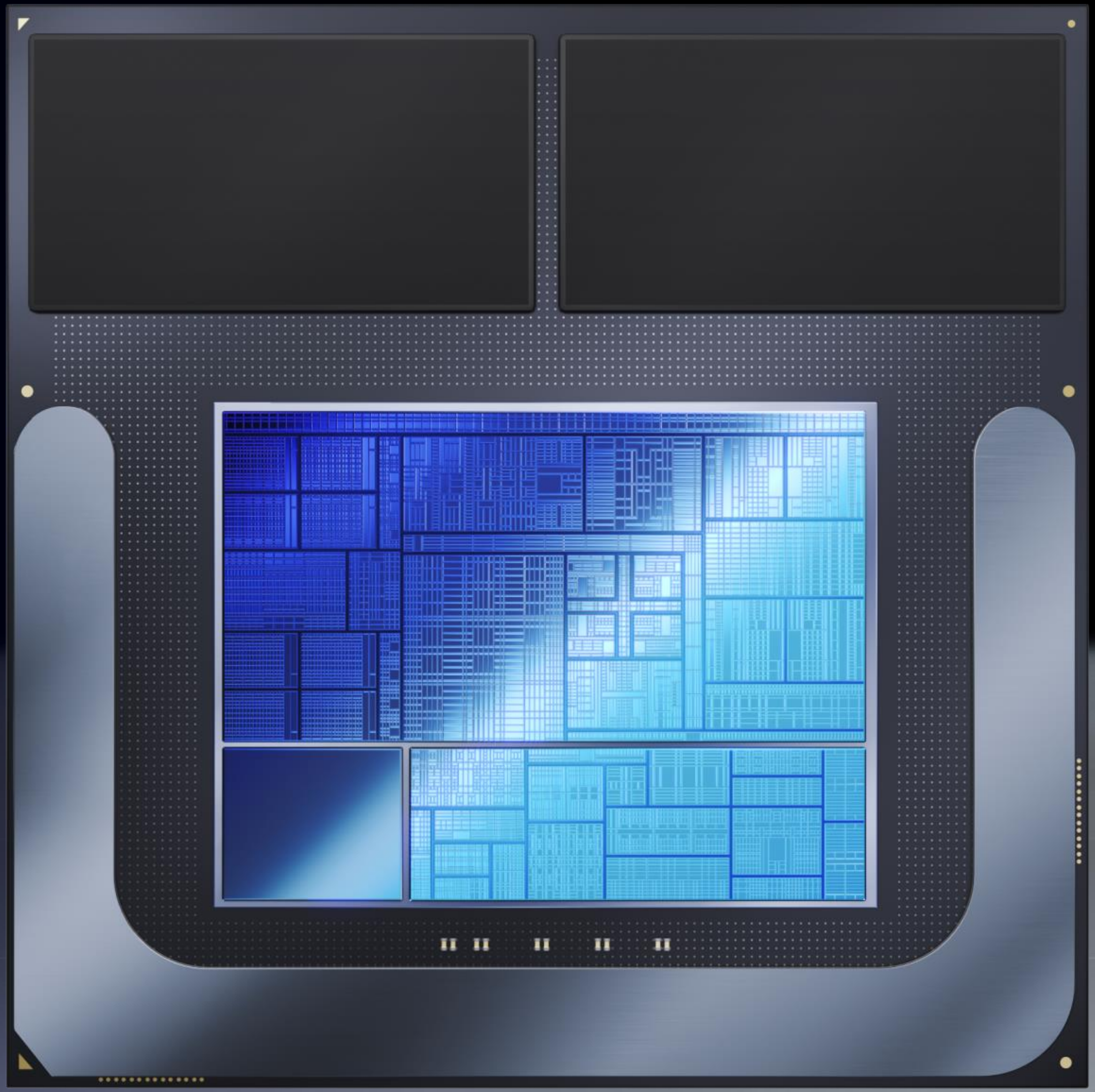
Up to
40%
lower PHY power

Up to
250mm^2
area savings



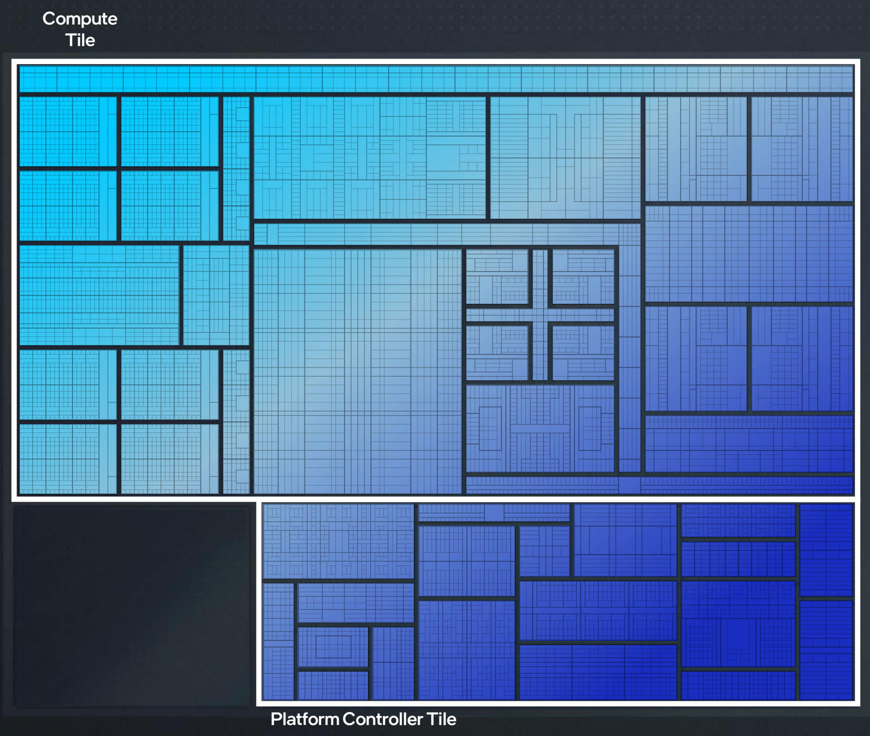
Lunar Lake

Architecture overview



Lunar Lake

Architectural Framework

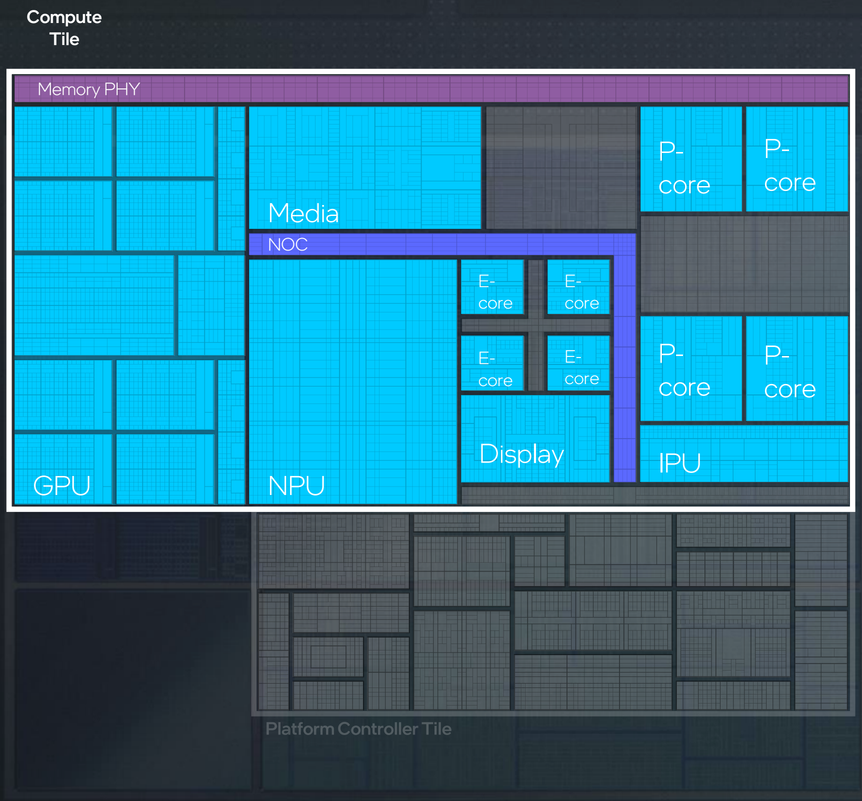


- Compute tile structure
- Memory side cache
- Enhanced E-core cluster
- New power delivery & management

Lunar
Lake

Enhanced SoC Structure

For better performance efficiency



Compute tile structure

Monolithic die, on a leading process

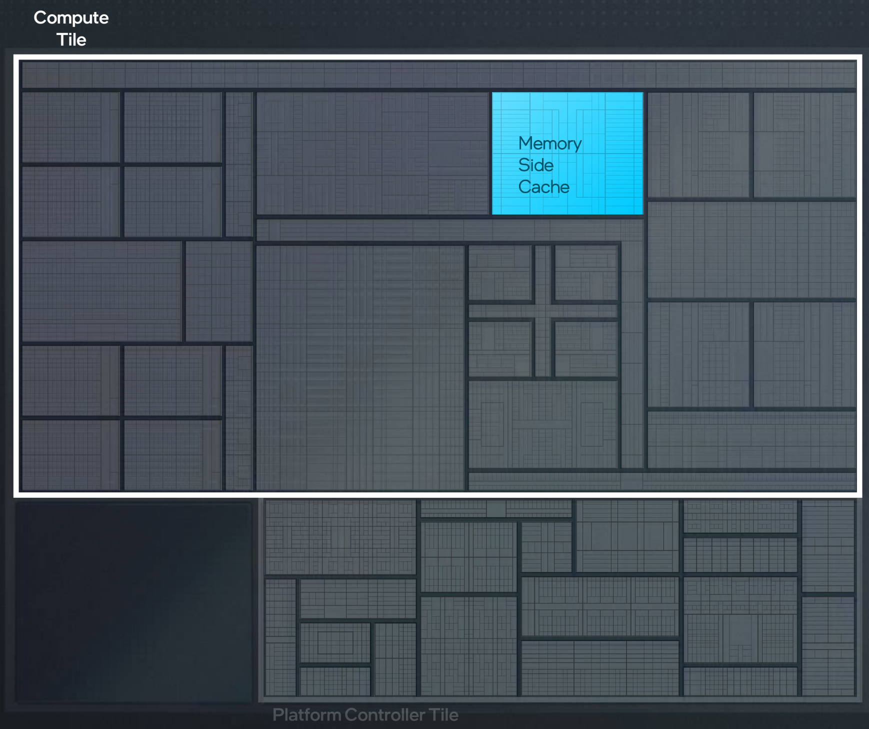
New NOC, with enhanced caching

Optimized memory latency

Lunar Lake

Memory Side Cache

For efficient performance & lower power consumption for other engines



Memory side cache

8MB physical cache

Reduce DRAM traffic and power

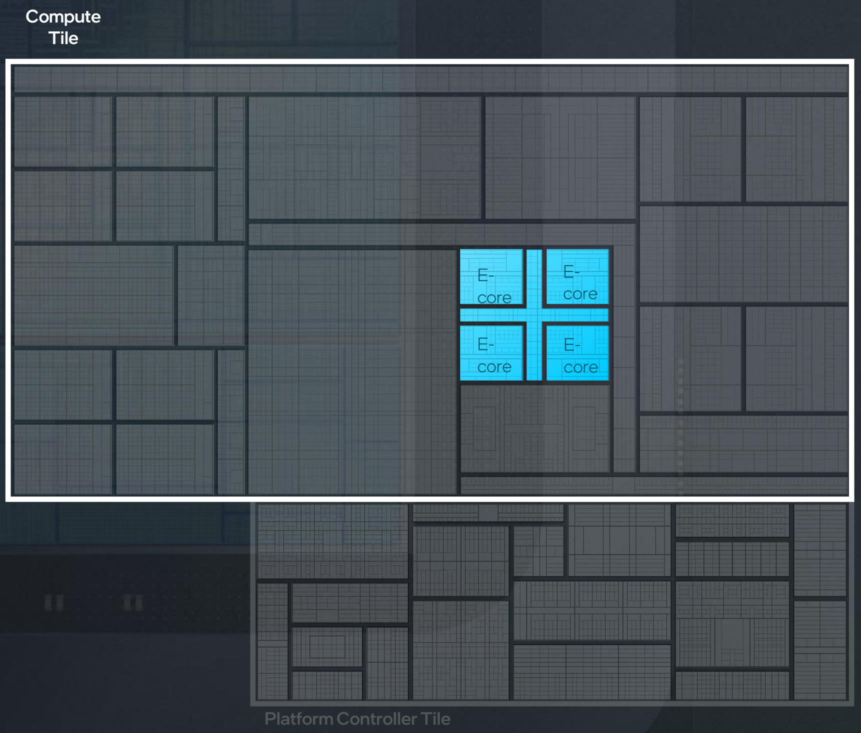
Enhanced both latency and bandwidth

Caching for IO engines

Lunar Lake

Enhanced E-core Cluster

For more efficient performance & lower power for almost every use[1]



Enhanced E-core cluster

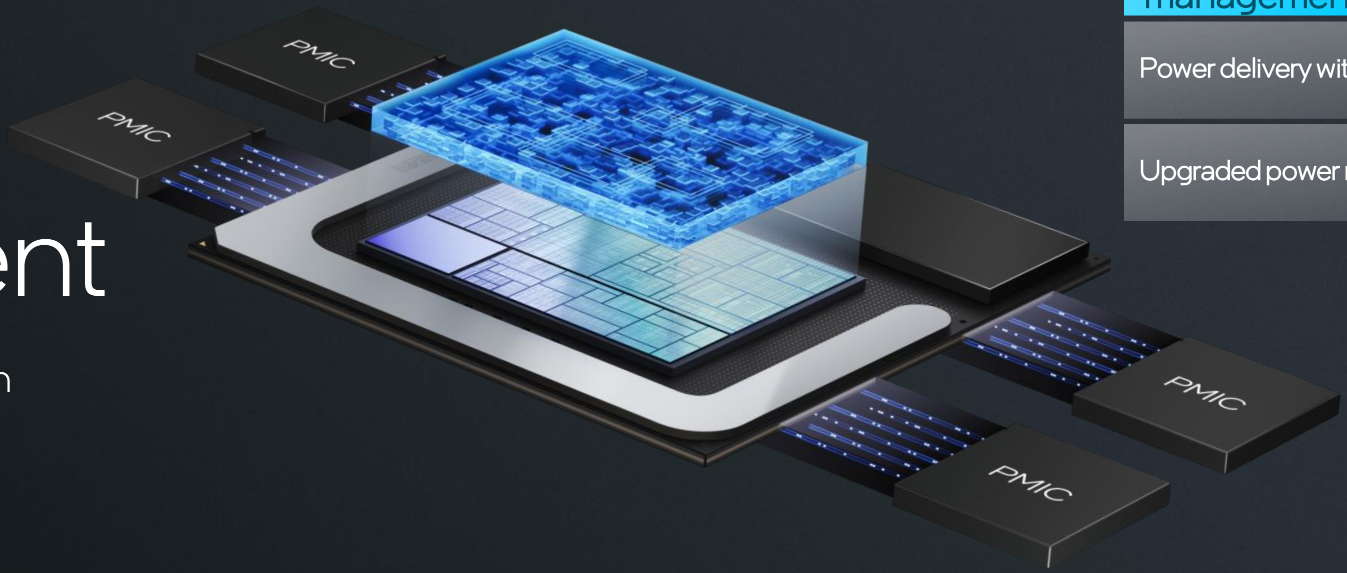
- Doubled the core count
- Increased 4MB L2 cache
- Leading-edge process for frequency and power
- Mem side cache for power and latency
- Independent power delivery

[1] vs. previous gen low power island

Lunar Lake

Power Delivery & Management

For optimal SOC power utilization & performance efficiency



New power delivery & management

Power delivery with 4 PMICs

Upgraded power management

Lunar Lake

Power Delivery & Management

For optimal SOC power utilization & performance efficiency

New power delivery & management

Power delivery with 4 PMICs

- More power rails
- Dynamic voltage ID
- More telemetry (IMON)

Lunar Lake

Power Delivery & Management

For optimal SOC power utilization & performance efficiency

New power delivery & management

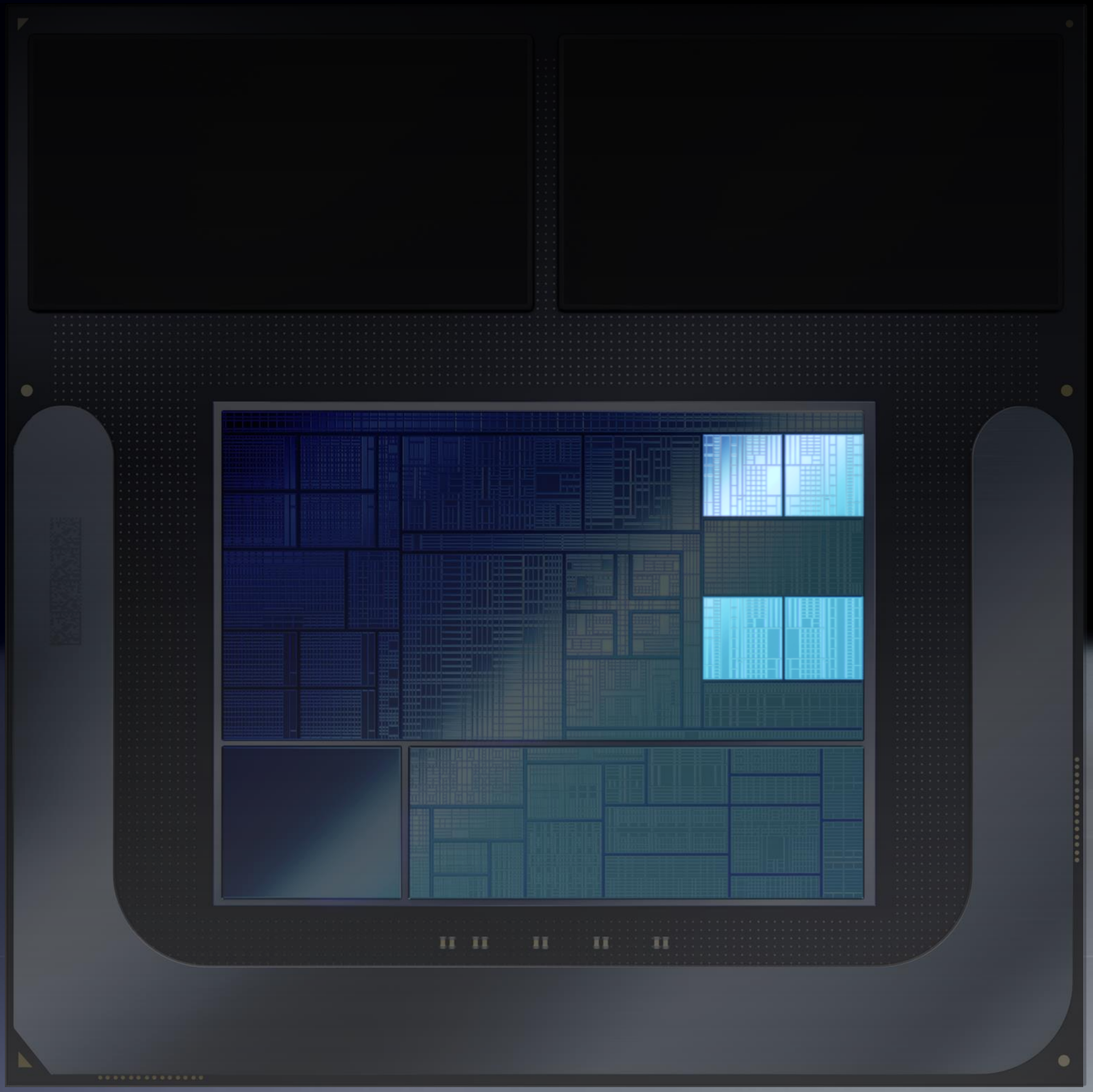
Power delivery with 4 PMICs

Upgraded power management

- Intel Thread Director focus on efficiency
- Optimized power balancer per load type
- Enhanced "sleep" states power and latency
- ML-based WL classification & frequency control

Lunar Lake

Performance cores



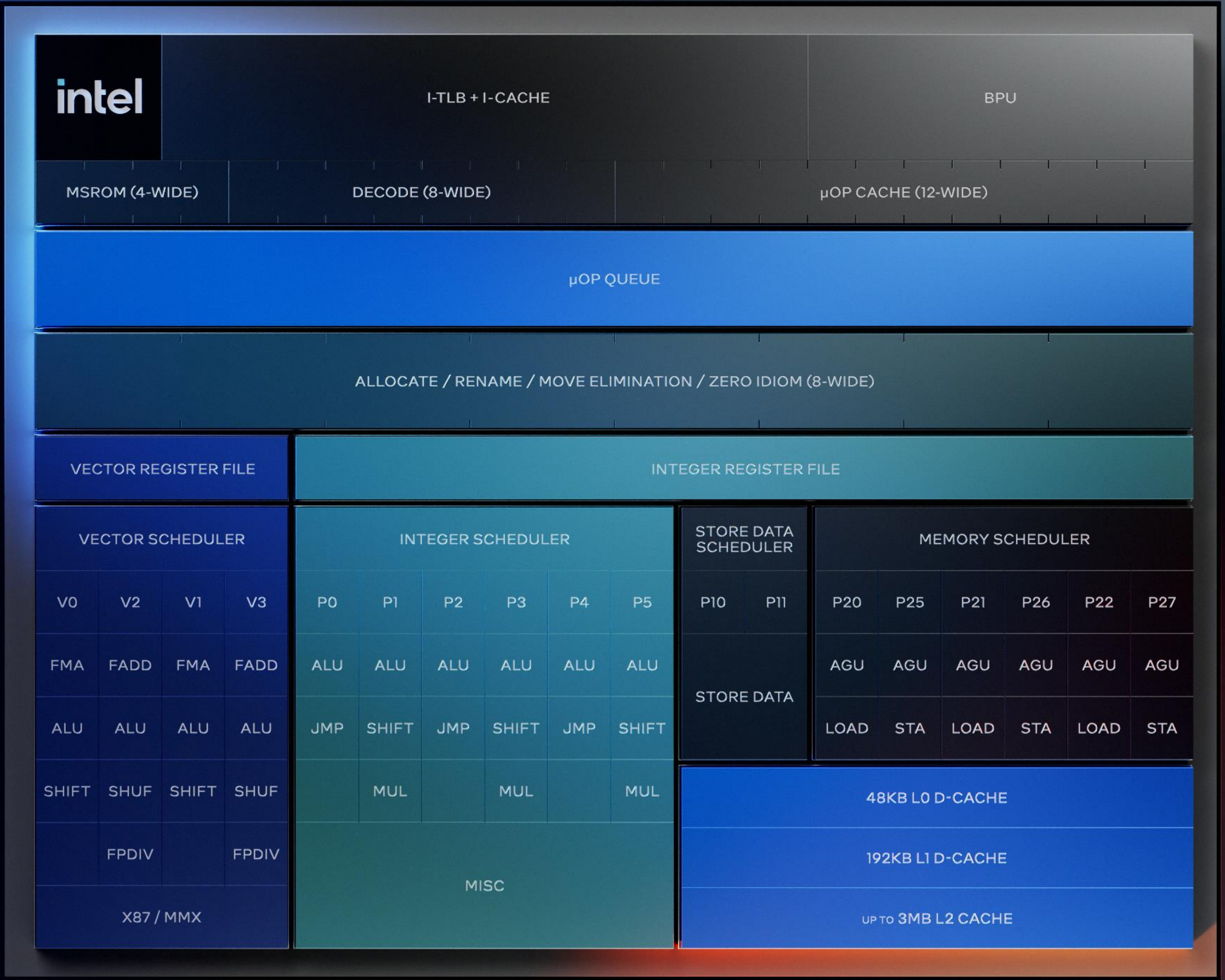
Lion Cove P-core

Architecture goals

Performance & area efficiency
Optimize ST perf/watt and perf/area for client SoCs

Overhaul microarchitecture
Generational IPC improvement and future scalability

Modernize design database
Accelerate innovation going forward



I-TLB + I-CACHE

BPU

MSROM (4-WIDE)

DECODE (8-WIDE)

µOP CACHE (12-WIDE)

µOP QUEUE

ALLOCATE / RENAME / MOVE ELIMINATION / ZERO IDIOM (8-WIDE)

VECTOR REGISTER FILE

INTEGER REGISTER FILE

VECTOR SCHEDULER

V0	V2	V1	V3
FMA	FADD	FMA	FADD
ALU	ALU	ALU	ALU
SHIFT	SHUF	SHIFT	SHUF
	FPDIV		FPDIV

X87 / MMX

INTEGER SCHEDULER

P0	P1	P2	P3	P4	P5
ALU	ALU	ALU	ALU	ALU	ALU
JMP	SHIFT	JMP	SHIFT	JMP	SHIFT
	MUL		MUL		MUL

MISC

STORE DATA SCHEDULER

P10	P11
STORE DATA	

MEMORY SCHEDULER

P20	P25	P21	P26	P22	P27
AGU	AGU	AGU	AGU	AGU	AGU
LOAD	STA	LOAD	STA	LOAD	STA

48KB L0 D-CACHE

192KB L1 D-CACHE

UP TO 3MB L2 CACHE

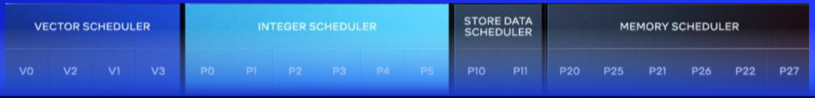
Lion Cove P-core

Rearchitected for efficient performance

AI-based

power management

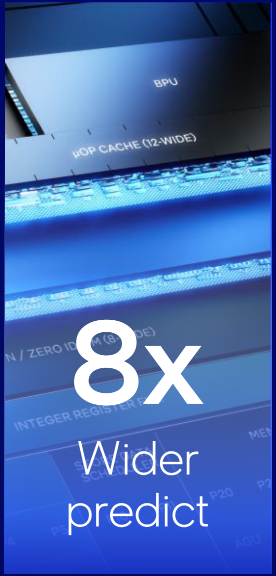
18 Execution ports



Up to 12MB

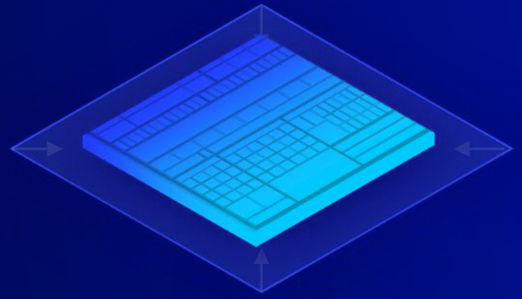
shared L3 cache

on Lunar Lake



8x

Wider predict



Optimized for

PPA

16.67MHz

Finer clock intervals

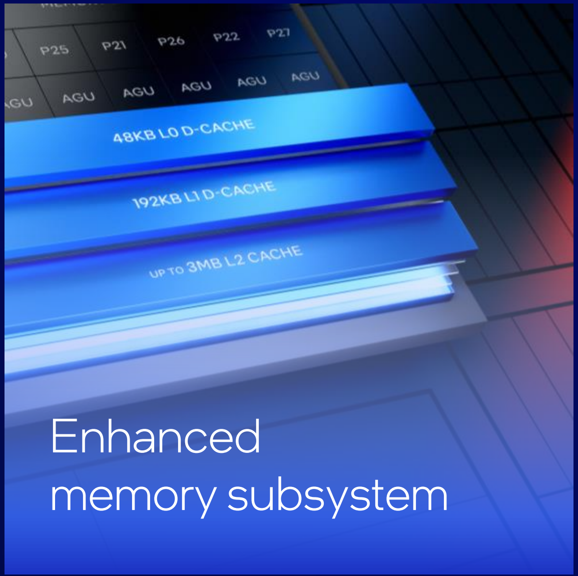
Wider scheduling

ALLOCATE / RENAME / MOVE ELIMINATION / ZERO IDIOM (8-WIDE)

across both allocation/rename & retire

VEC ◀ **Split** ▶ INT

out of order engine



Enhanced memory subsystem

intel.

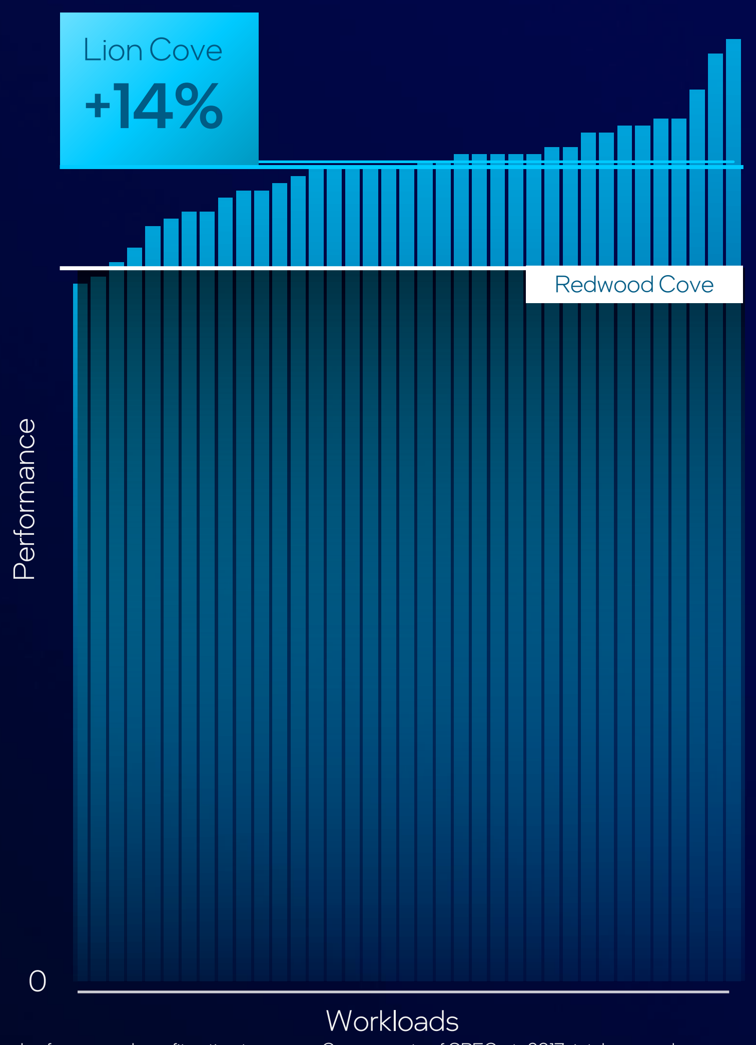
Lion Cove P-core

Double-digit performance gains over prior generation

🟦 Lion Cove in Lunar Lake

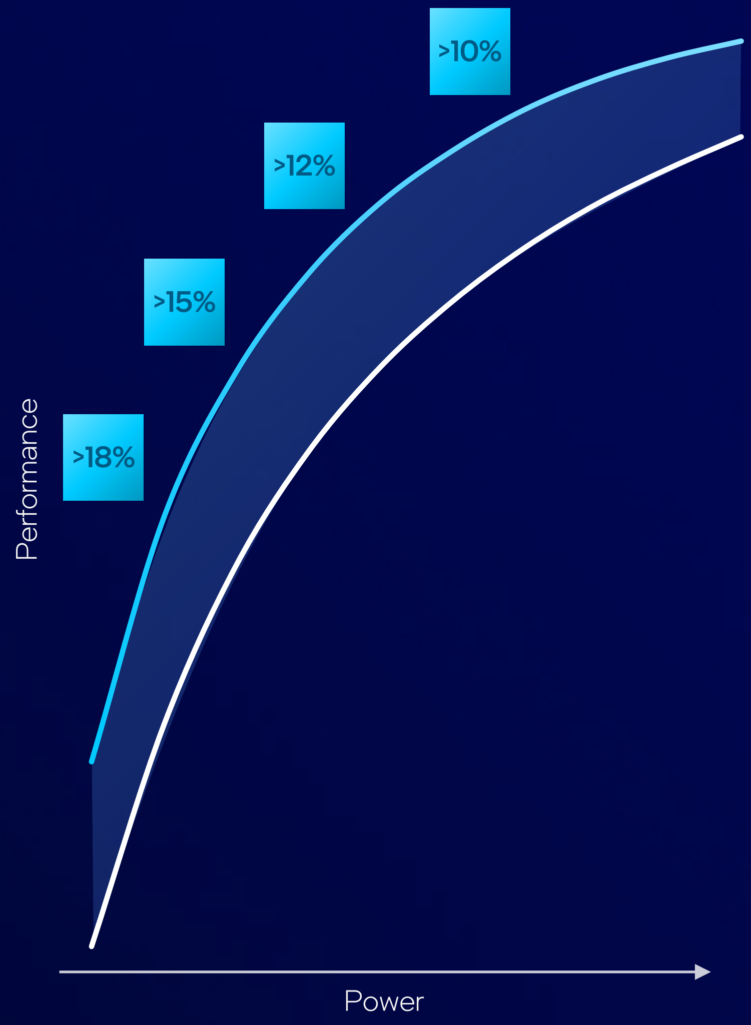
⬜ Redwood Cove in Meteor Lake

IPC



Iso frequency benefit estimate across: Components of SPECrate2017_int_base and SPECrate2017_fp_base (both estimated) running 1 copy, Cinebench R23 Single Core, Cinebench 2024 Single Core, Geekbench5.4.5 Single-Core, Geekbench6.2.1 Single-Core, WebXPRT 4, Speedometer

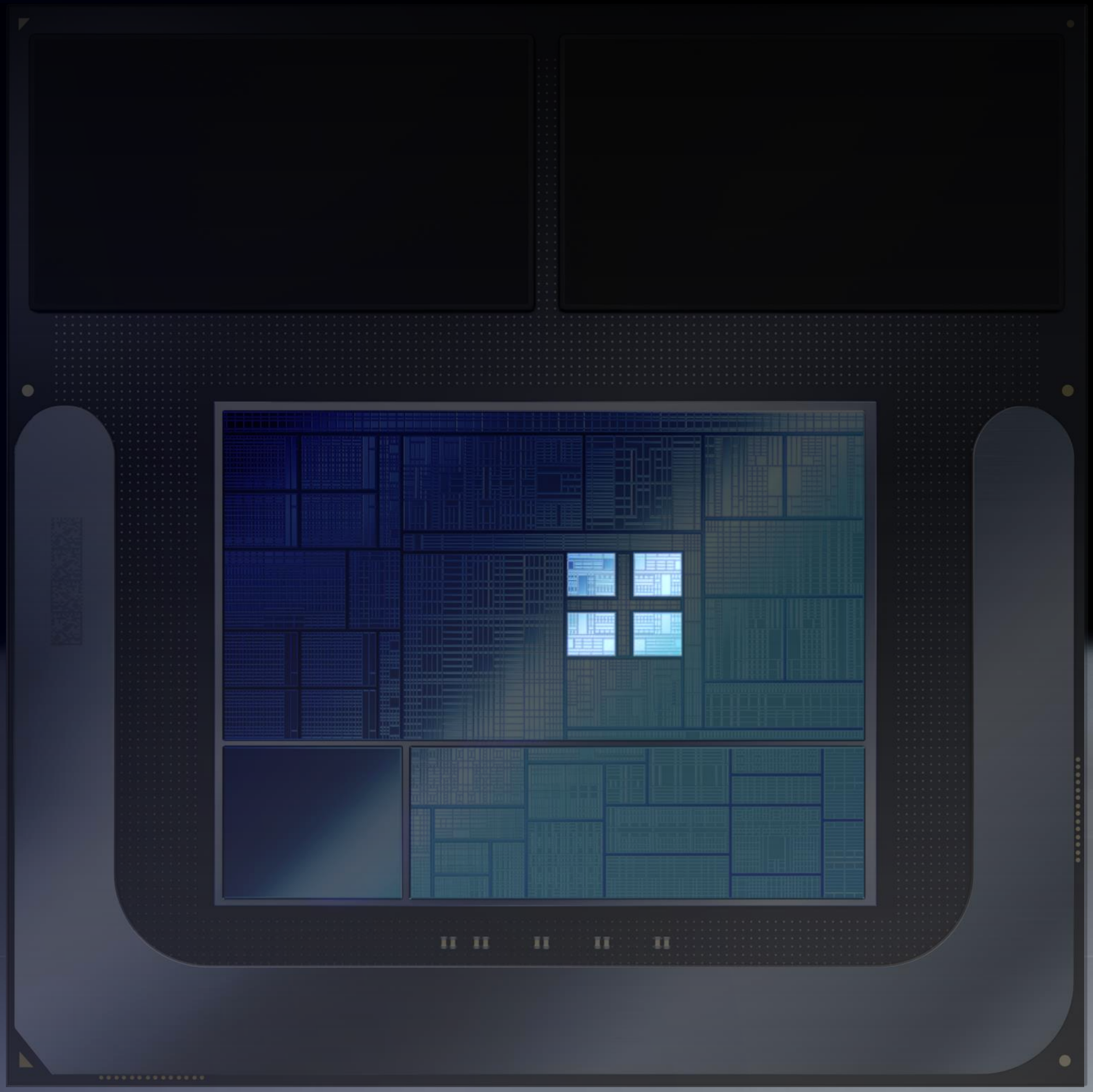
Performance at power



Results are based on SPECrate2017_int_base (estimated) running n copies. Based on measurement on an Intel internal reference validation platforms at a fixed PL1 power

Lunar Lake

Efficient cores



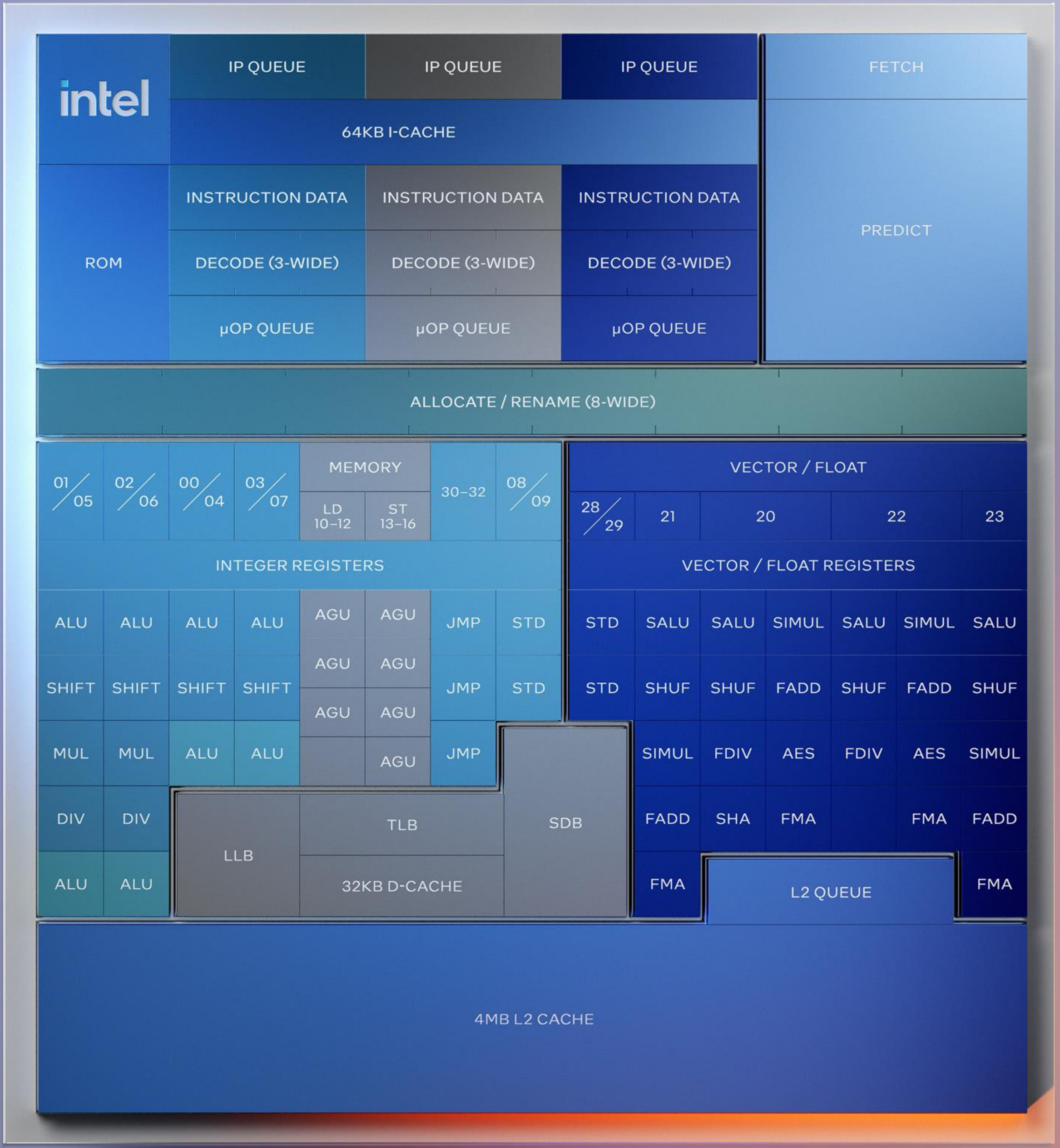
Skymont E-core

Architecture goals

Increase workload coverage
Increase range of low power island & MT perf

Double vector & AI throughput
For increased VNNI capability support

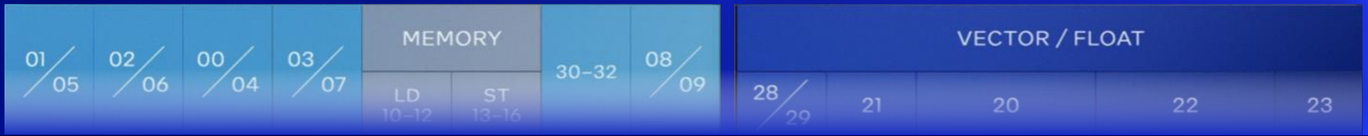
Increase scalability
For overall performance uplift



Skymont E-core

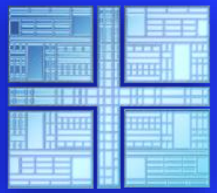
Our most efficient performant architecture

26 dispatch ports



Scalable & flexible
across multiple implementations

2x
L2 cache bandwidth



4MB
shared L2 cache

Enhanced prediction
to find instructions faster

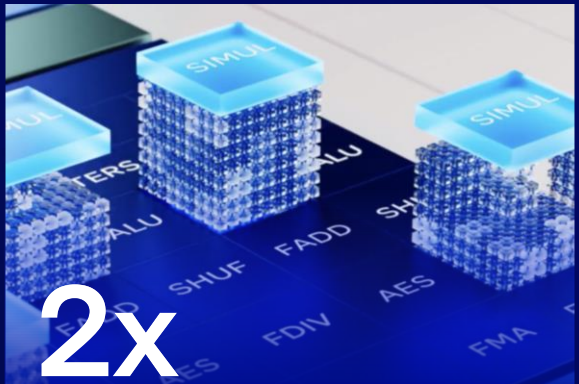
96B
parallel fetching

Deeper queueing
for better parallelism

Wider allocation & retire

ALLOCATE

RETIRE



2x
AI throughput
from 4x 128bit
FP &SIMD vector

intel.

Skymont IPC Gain

On Lunar Lake Low Power Island

vs. Meteor lake LP E-core

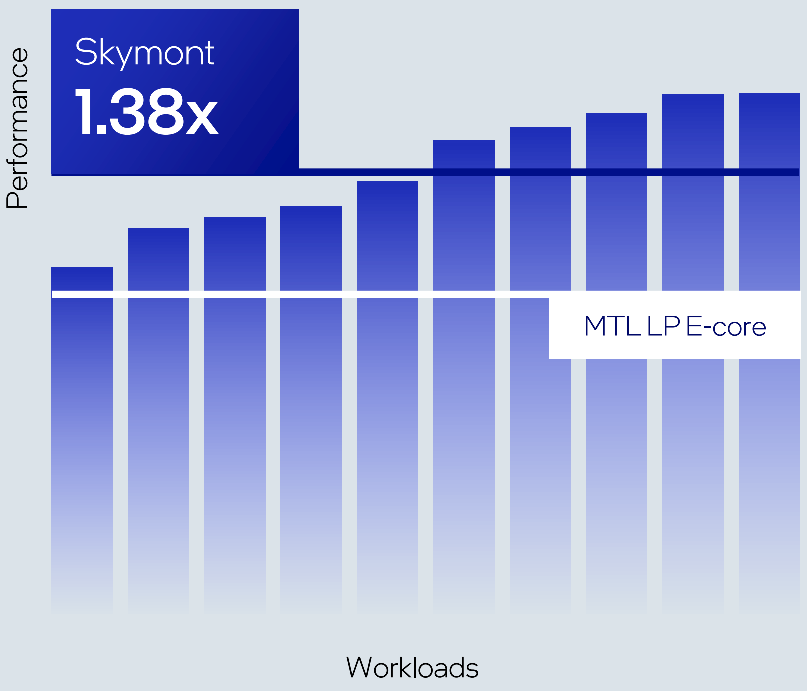
Skymont plus **Lunar Lake** fabric improvements combine to bring significant IPC improvements for many general workloads

SPEC CPU 2017 Rate
GCC12.1 -O2 Linux
Fixed Frequency (iso)

Results are based on Intel's internal projections/estimates (+/- 10% margin of error). See Intel.com/PerformanceIndex for additional details.

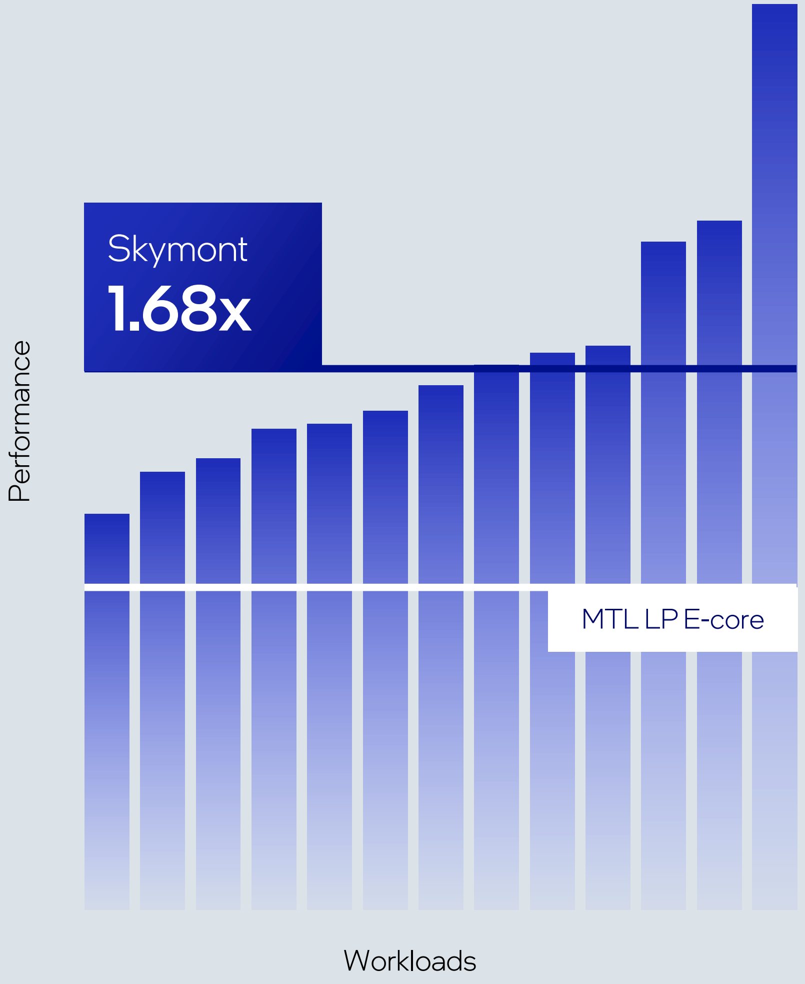
ST integer improvement

SPECrate2017_int_base est / GCC



ST FP improvement

SPECrate2017_fp_base est/ GCC



Skymont Power & Performance

On Lunar Lake Low Power Island

vs. Meteor lake LP E-core

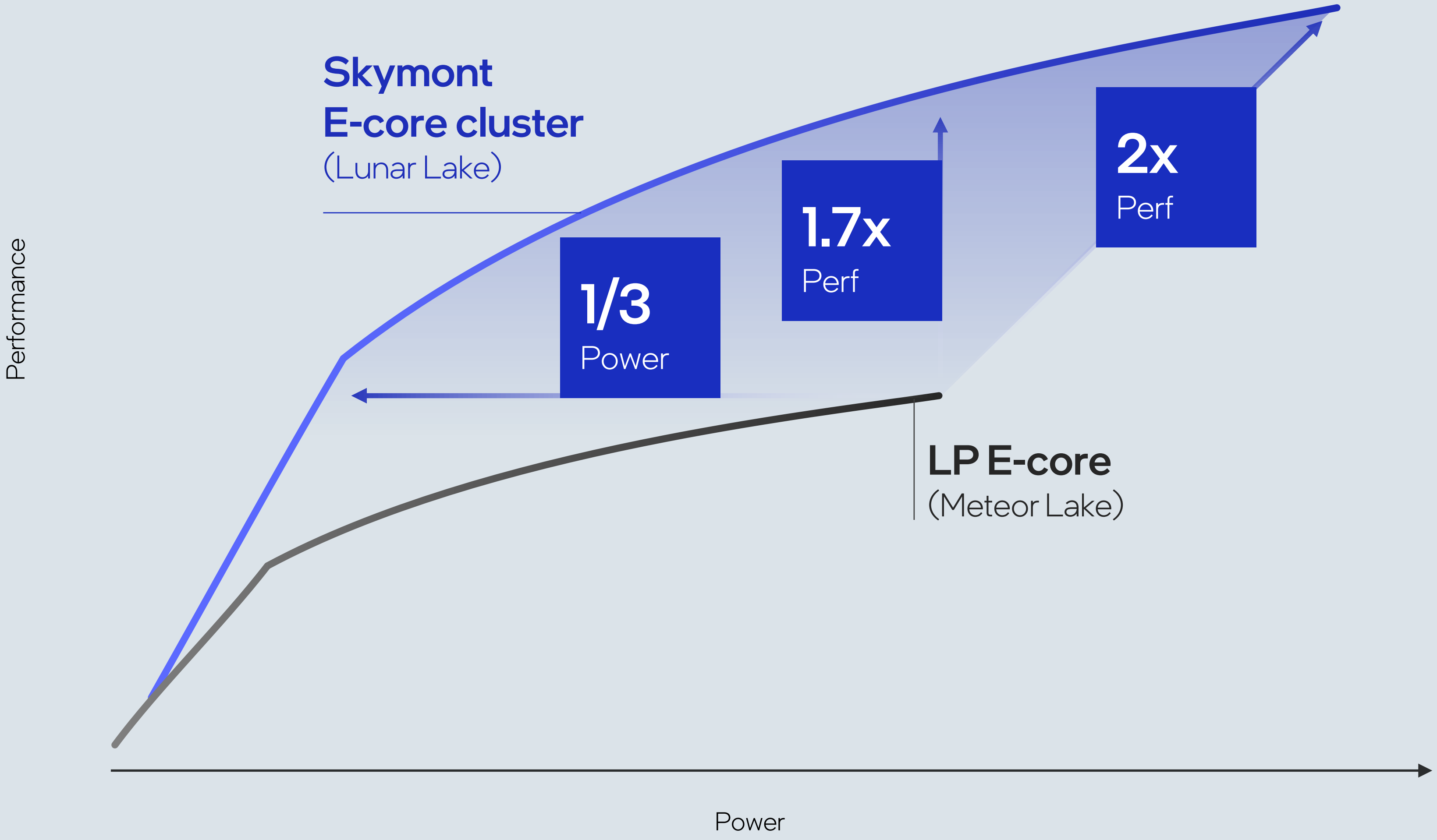
Significant increase in workload coverage for the Lunar Lake Low Power Island

More IPC, frequency & cores

SPECrate2017_int_base est
GCC12.1 -O2 Linux (iso)

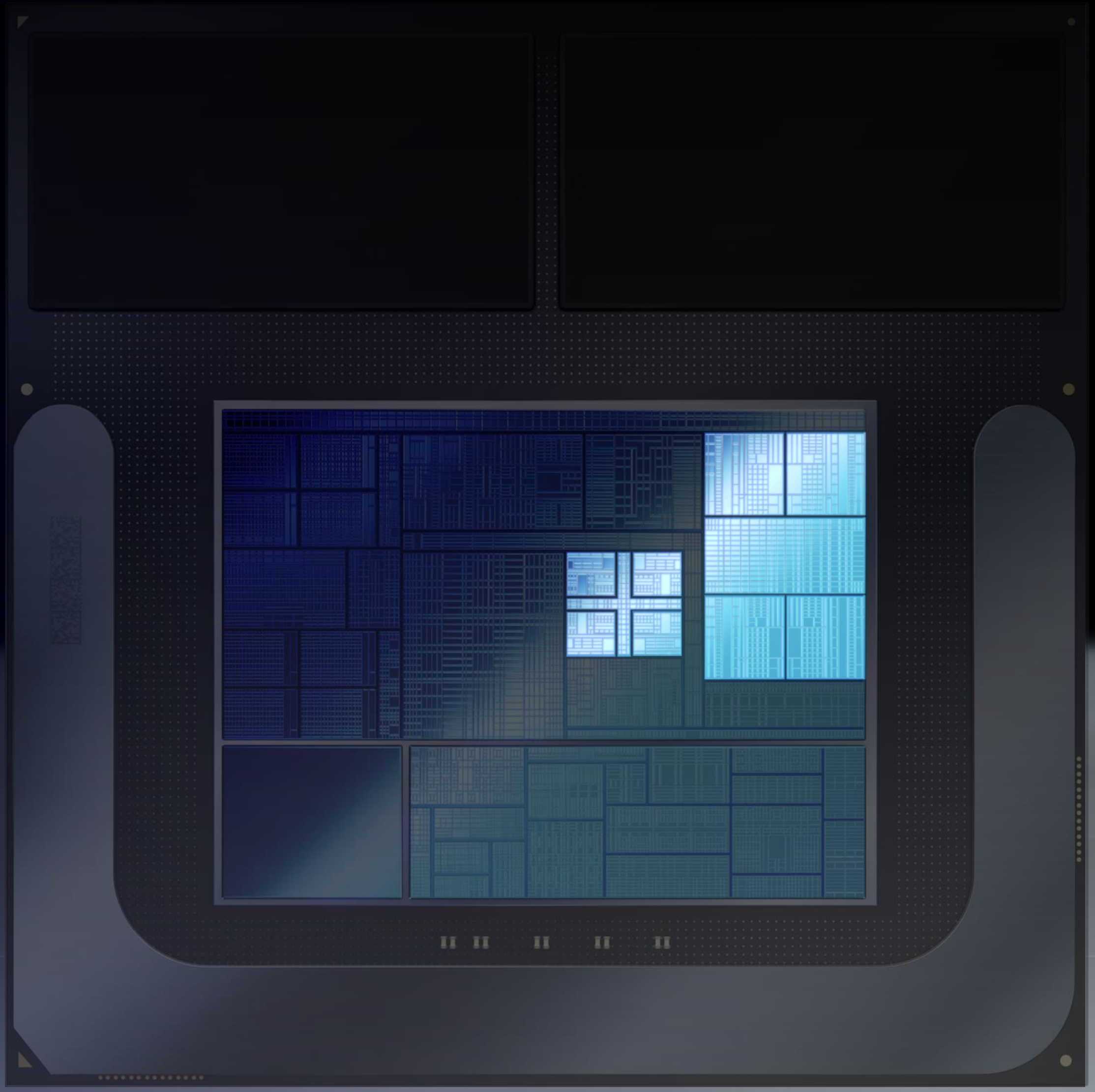
Results are based on SPECrate2017_int_base (estimated) running n copies. Based on measurement on an Intel internal reference validation platforms at a fixed PL1 power

Single threaded

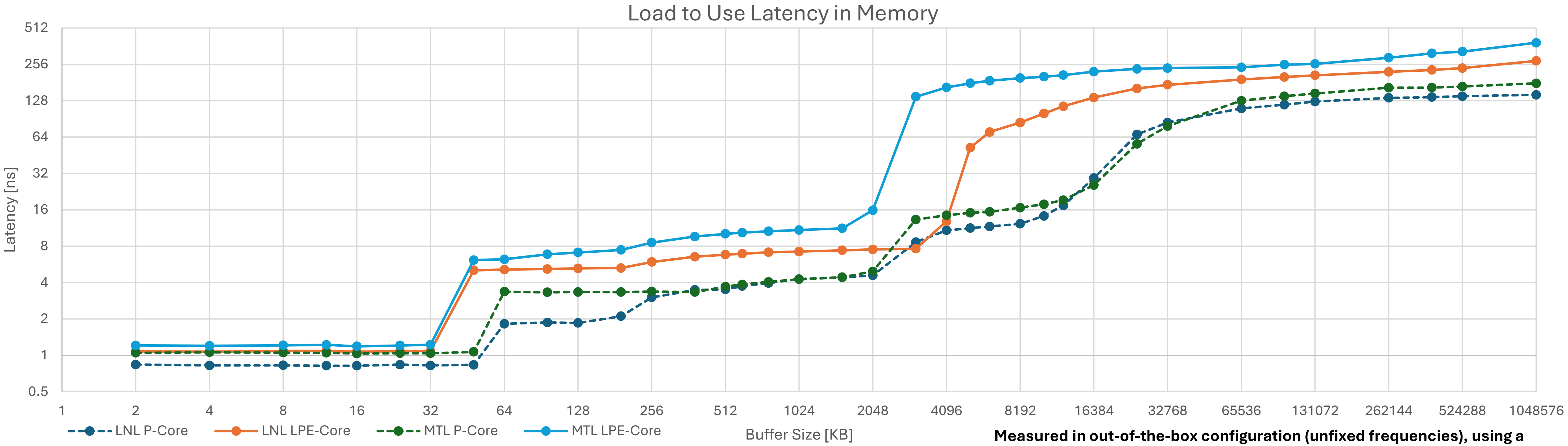


Lunar Lake

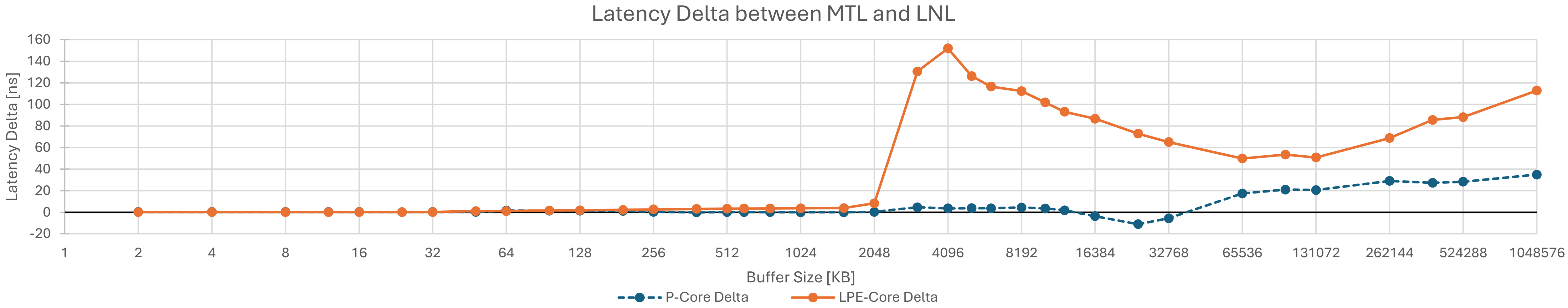
E-core cluster & P-core cluster



Latency Optimizations

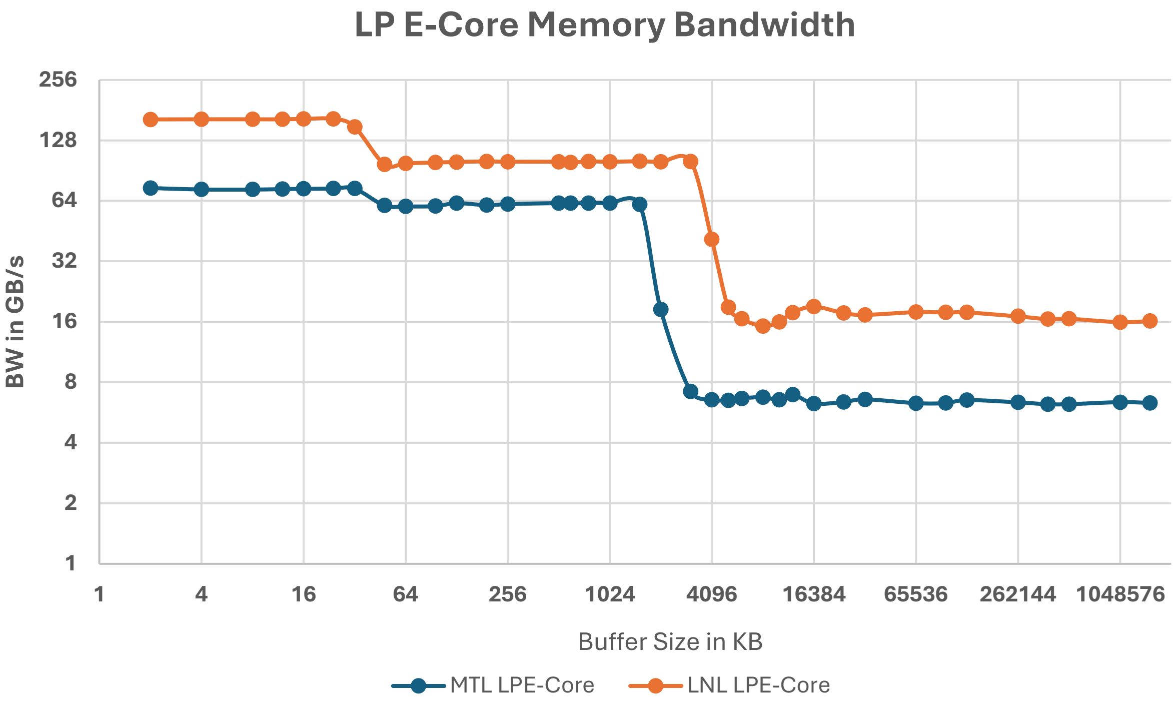


Measured in out-of-the-box configuration (unfixed frequencies), using a pointer chasing latency test with 4KB Page size on Windows 11 at 15W-17W



Efficient Cores BW & Cross Cluster latencies

2.8x Memory Bandwidth to LP Efficient Cluster



As part of the capable efficient Cores cluster that can keep the Compute Cluster powered off more often

Low Latency Cross-Cluster Coherency

	Compute Cluster Cores				Low Power Island Cores			
Compute Cluster Cores		24.8	24.65	26	54.35	55.05	54.7	54.3
	23.85		27.25	27.45	54.65	55.2	55.6	54.1
	25.25	27.05		28.2	55.2	55.15	55.3	55.2
	25	27.5	27.9		56	55.15	55.3	56
Low Power Island Cores	55.05	54.45	55.4	55		23.7	23.6	22.7
	54.55	54.85	55.5	55.35	22.75		23.7	23.2
	55.1	55.05	55.55	55.15	23.4	23.6		23.5
	55.15	55.1	54.9	56.3	22.9	22.95	23.9	

~55 ns on LNL , 3X better vs Prev Gen

Low overhead for shared data in MT Workloads in both clusters enables efficient Cores cluster to also accelerate Multi-Threaded applications.

Breakthrough x86 Efficiency

Whilst covering the full CPU performance range

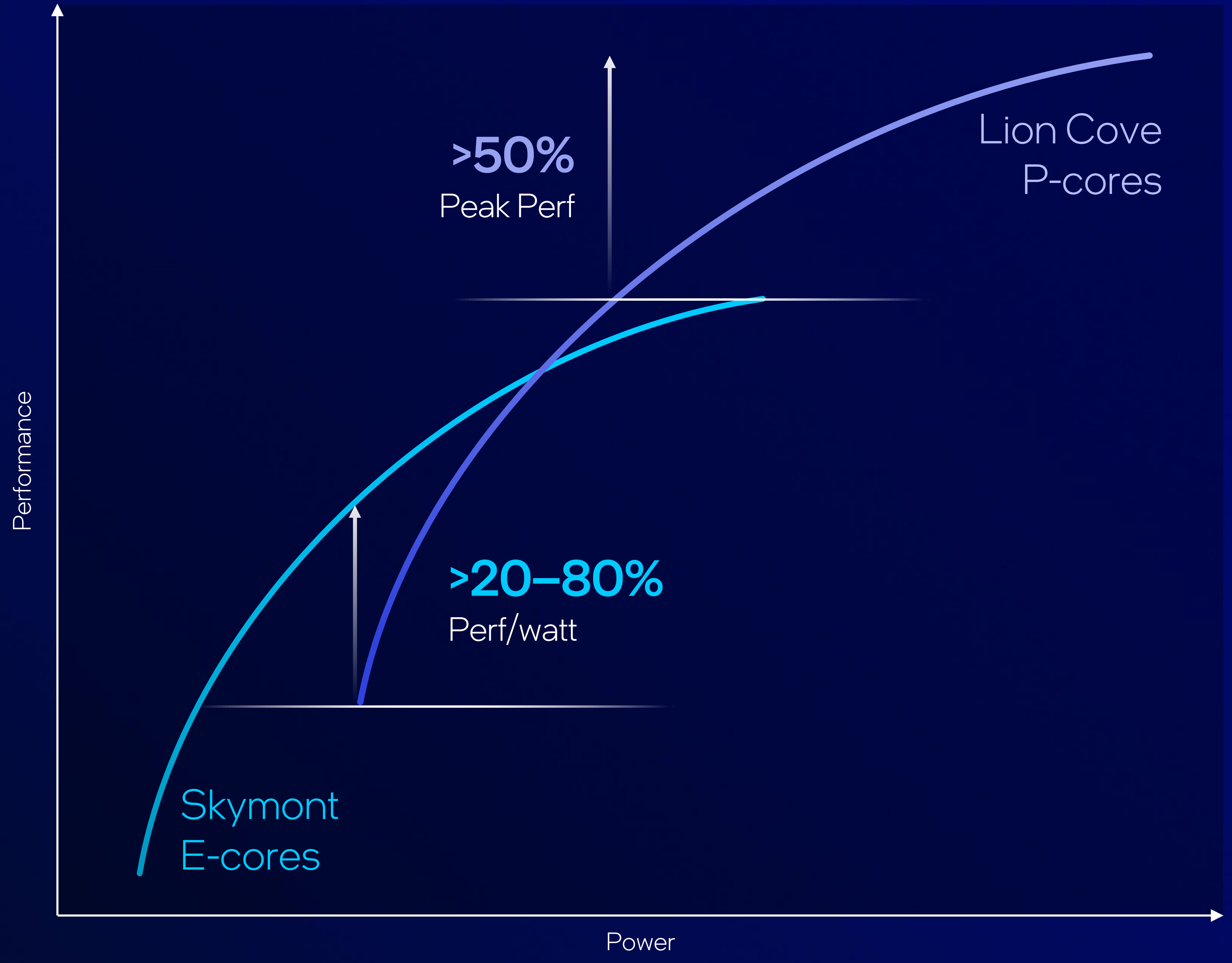
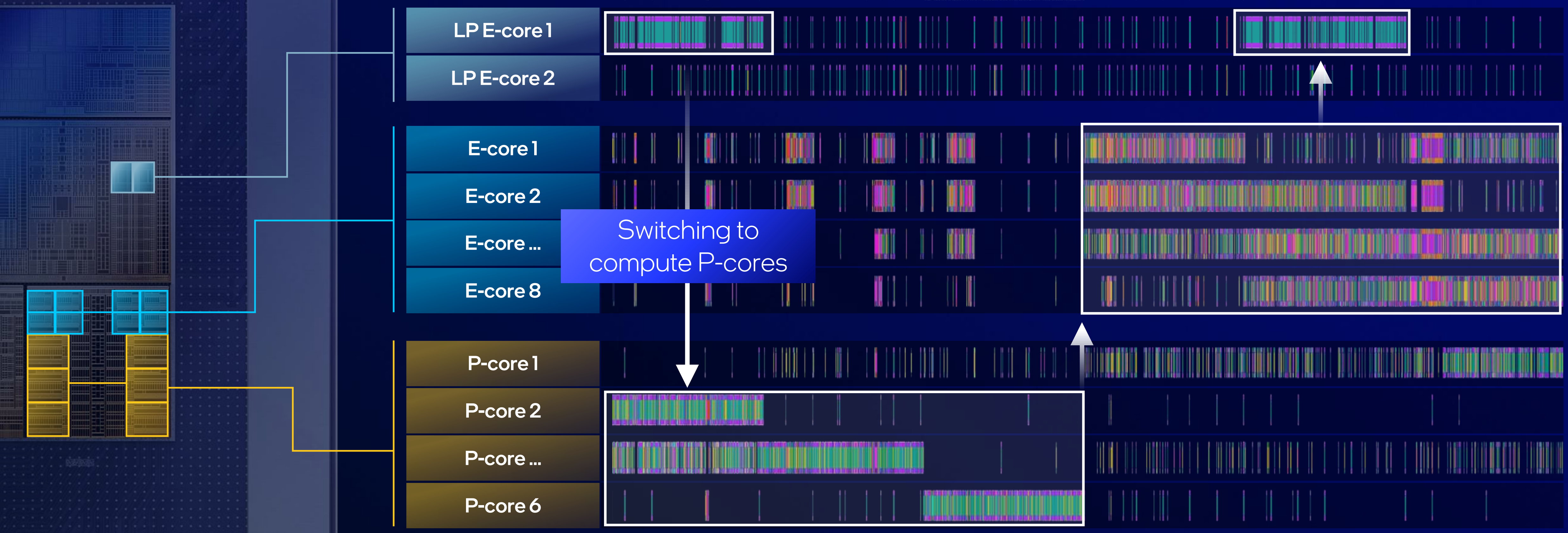


Illustration of the relative Lunar Lake P-core and E-core performance across the SoC power range.
See Intel.com/PerformanceIndex for more details.

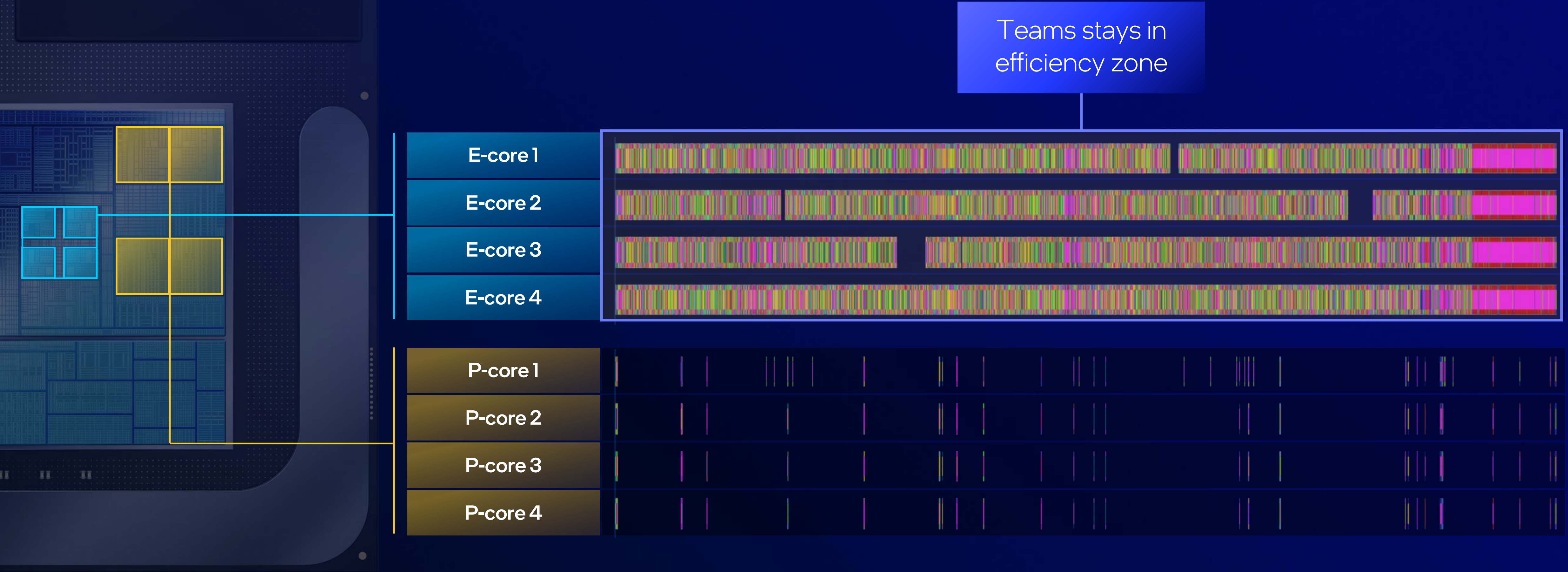
Teams Power Efficiency

Meteor Lake Scheduling



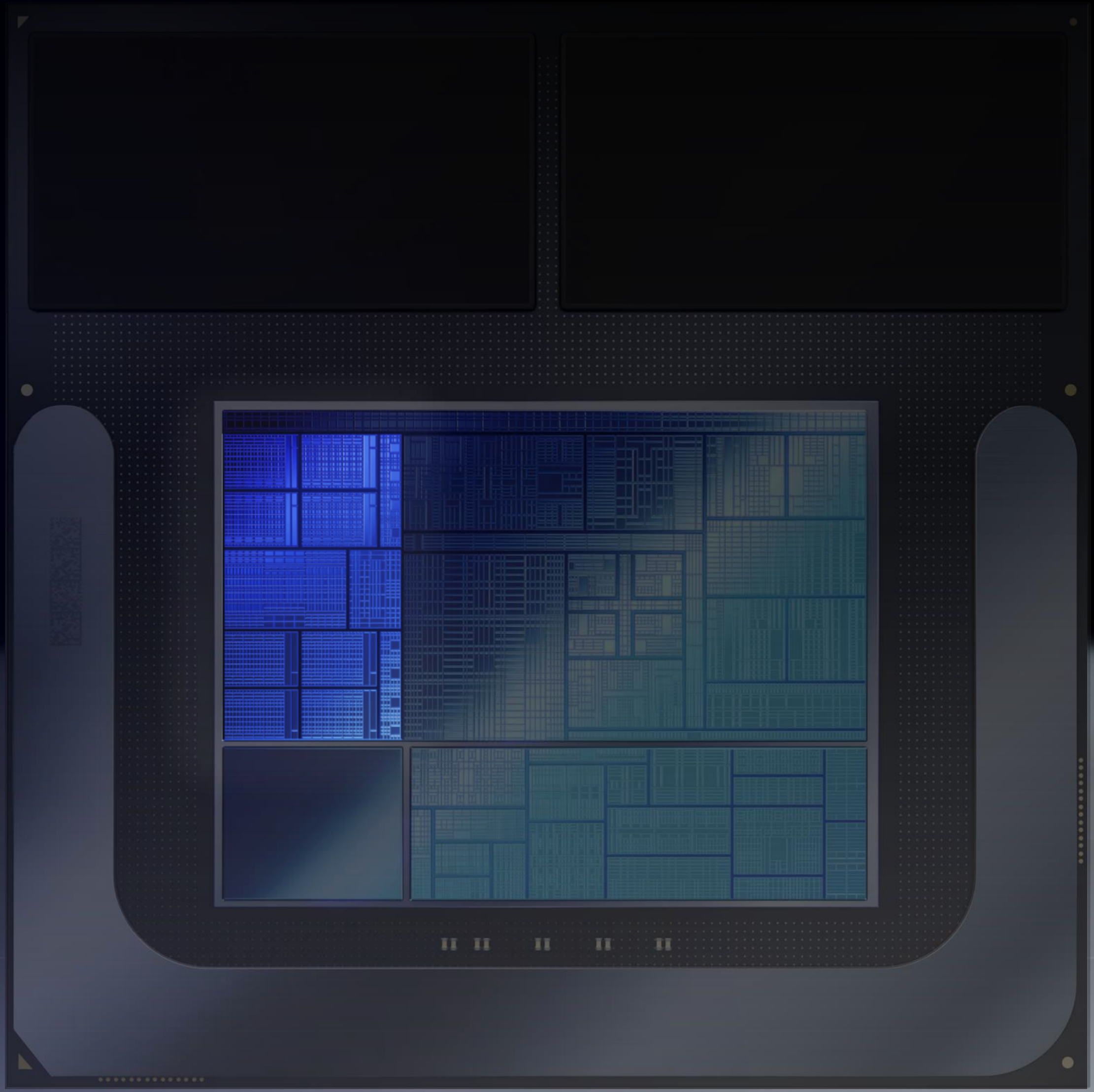
Teams Power Efficiency

Lunar Lake Scheduling



Lunar Lake

New X^e 2 GPU



Next Gen

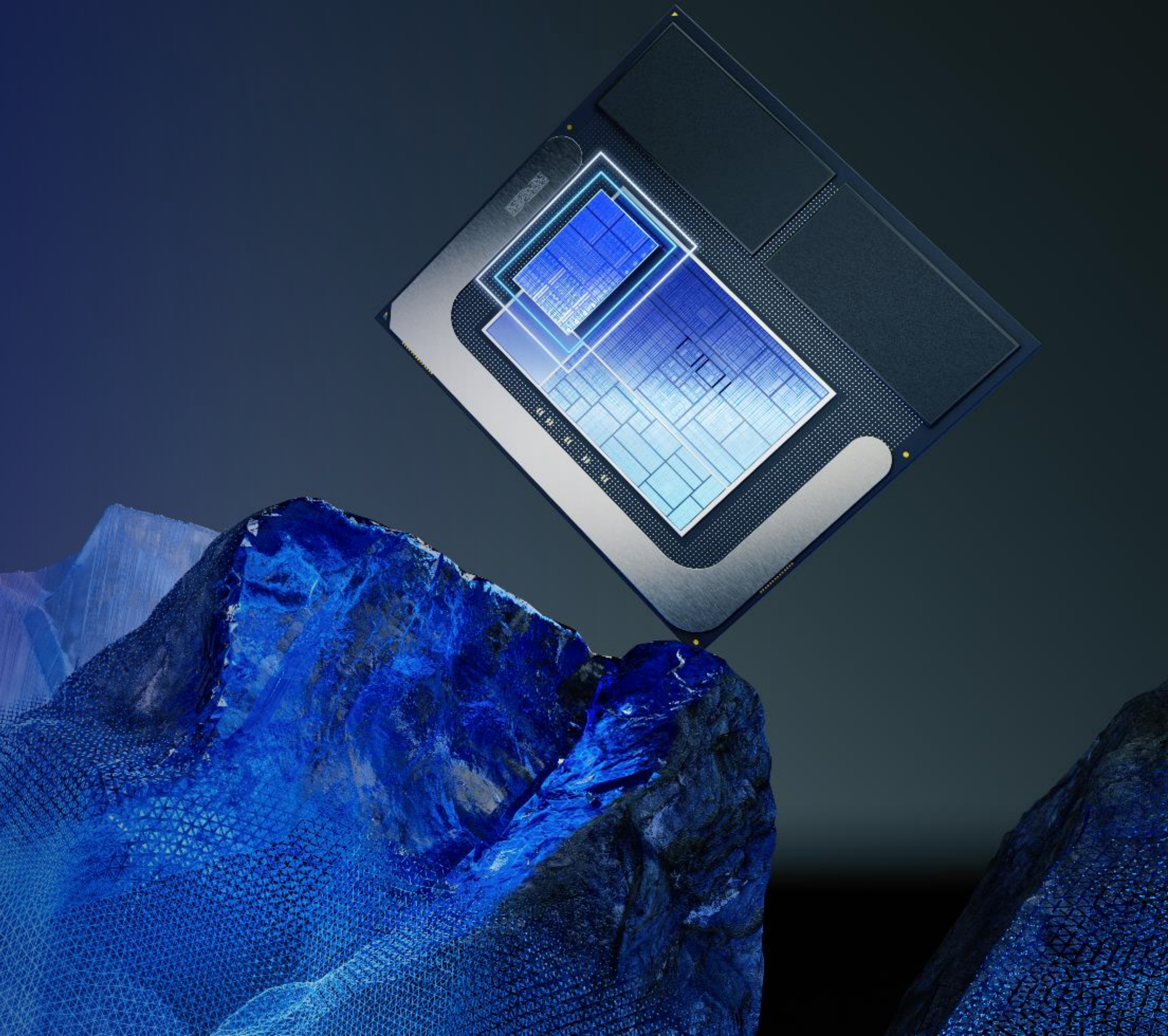
X^e2 GPU

Architecture goals

Improved utilization
of hardware functions

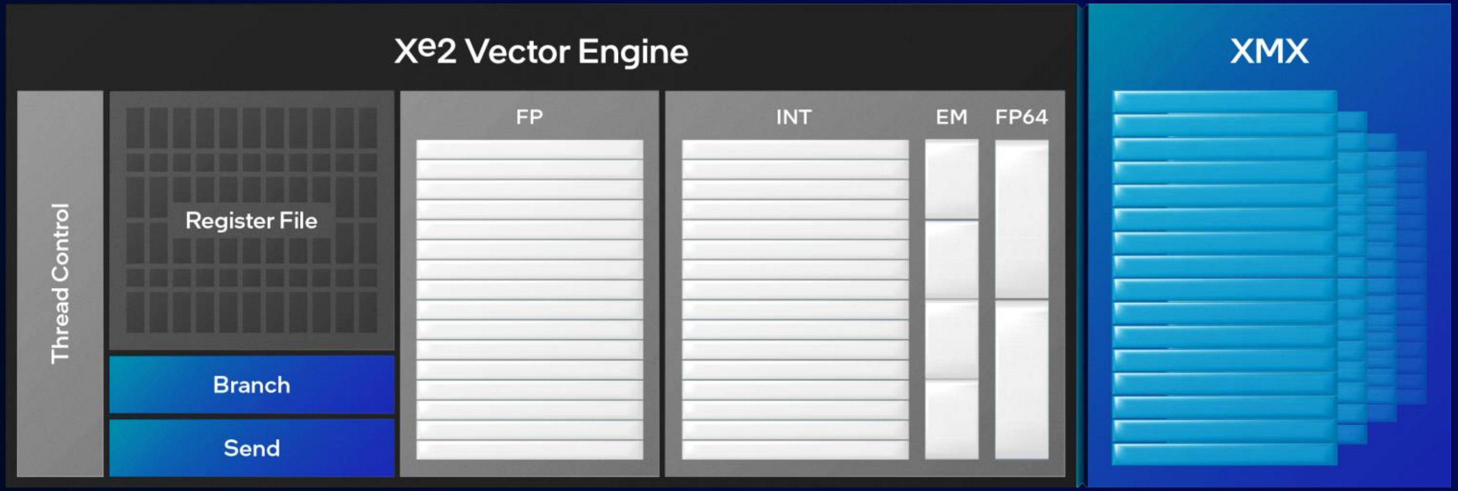
Improved distribution
of workload across architecture

Improved integration
Of hardware & software



New Vector Engine

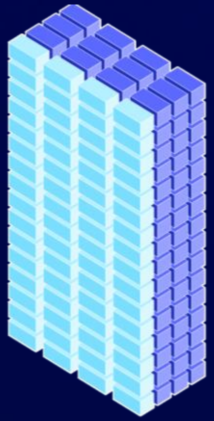
Optimized for efficiency & AI throughput



New Xe Matrix Extension Engines

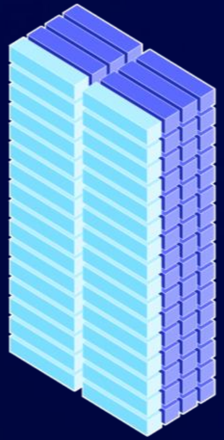
INT8

4096 OPS/clock



FP16

2048 OPS/clock

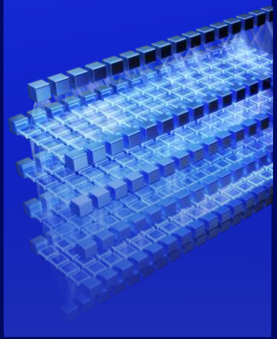


Next Gen Xe2 GPU

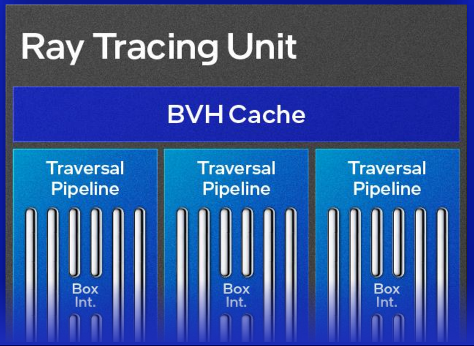
Major leap in graphics performance

up to 67 TOPS

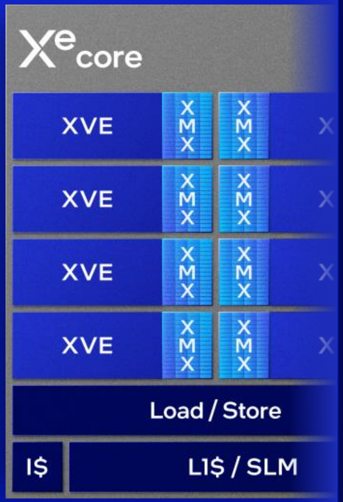
New XMX engines



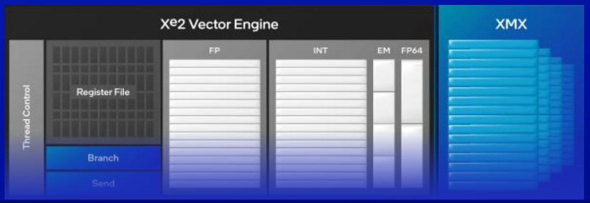
8 Larger ray tracing units



8 2nd gen Xe cores



Xe2 vector engines

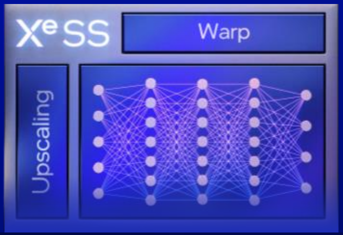


1.5x better vs. Meteor Lake GPU

intel ARC™ Software stack

eDP 1.5

Enhanced XeSS kernels



8 MB L2 cache

intel.

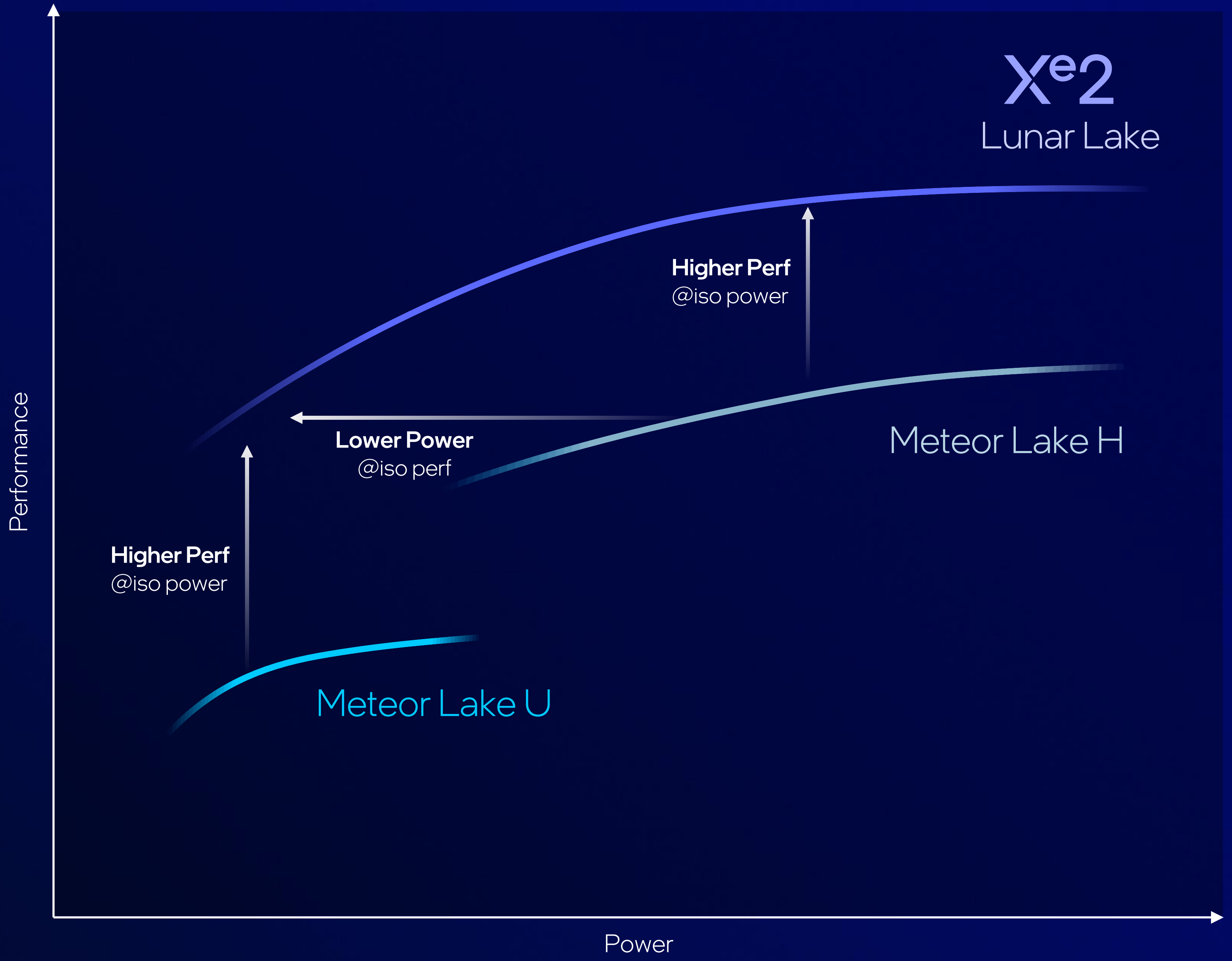
Next Gen

Xe2 GPU

Major leap in graphics performance

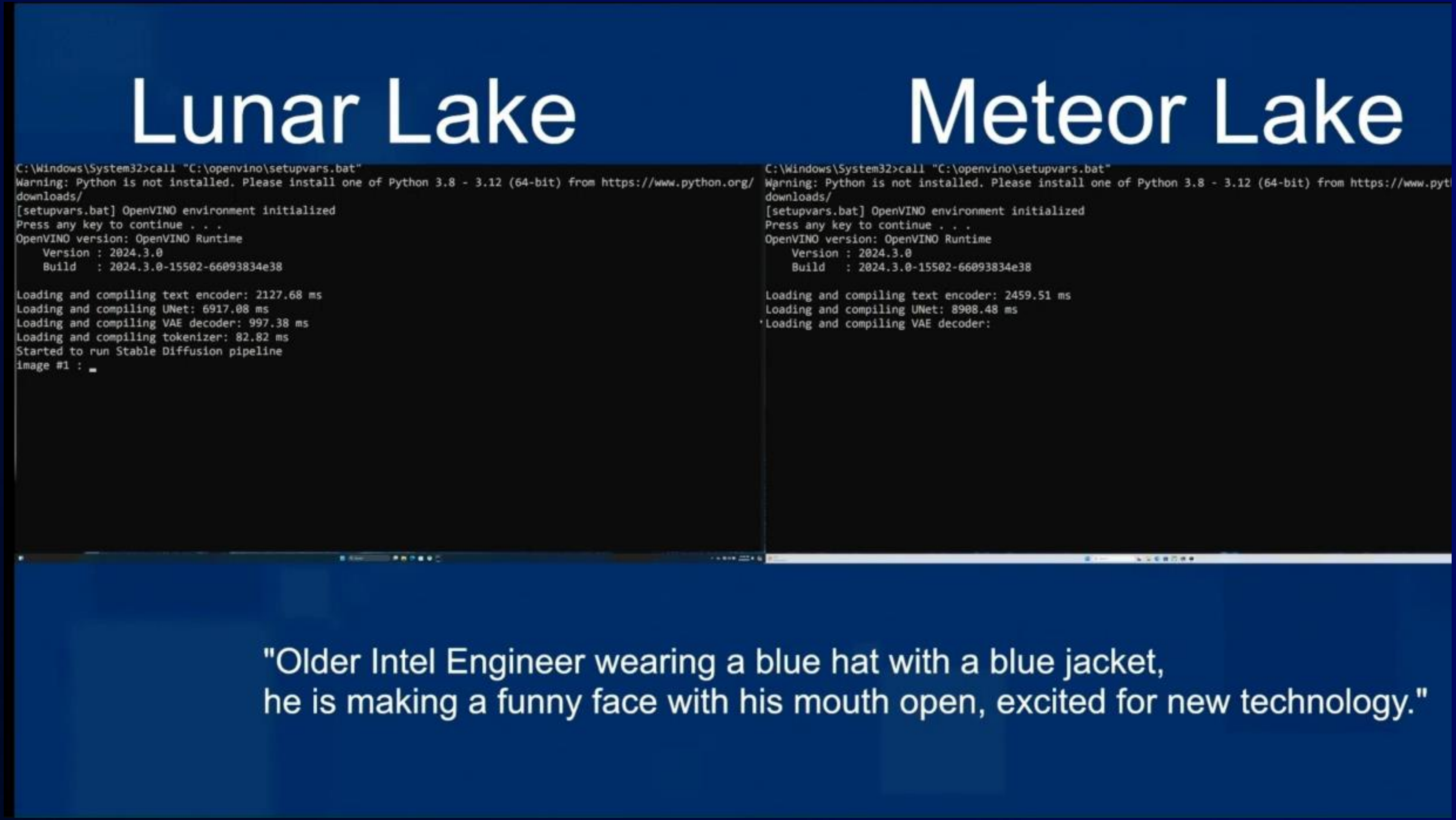
~1.5x

vs. previous gen



See backup for workloads and configurations. Results may vary.

Stable Diffusion



```
C:\Windows\System32>call "C:\openvino\setupvars.bat"
Warning: Python is not installed. Please install one of Python 3.8 - 3.12 (64-bit) from https://www.python.org/
downloads/
[setupvars.bat] OpenVINO environment initialized
Press any key to continue . . .
OpenVINO version: OpenVINO Runtime
    Version : 2024.3.0
    Build   : 2024.3.0-15502-66093834e38

Loading and compiling text encoder: 2127.68 ms
Loading and compiling UNet: 6917.08 ms
Loading and compiling VAE decoder: 997.38 ms
Loading and compiling tokenizer: 82.82 ms
Started to run Stable Diffusion pipeline
image #1 : _
```

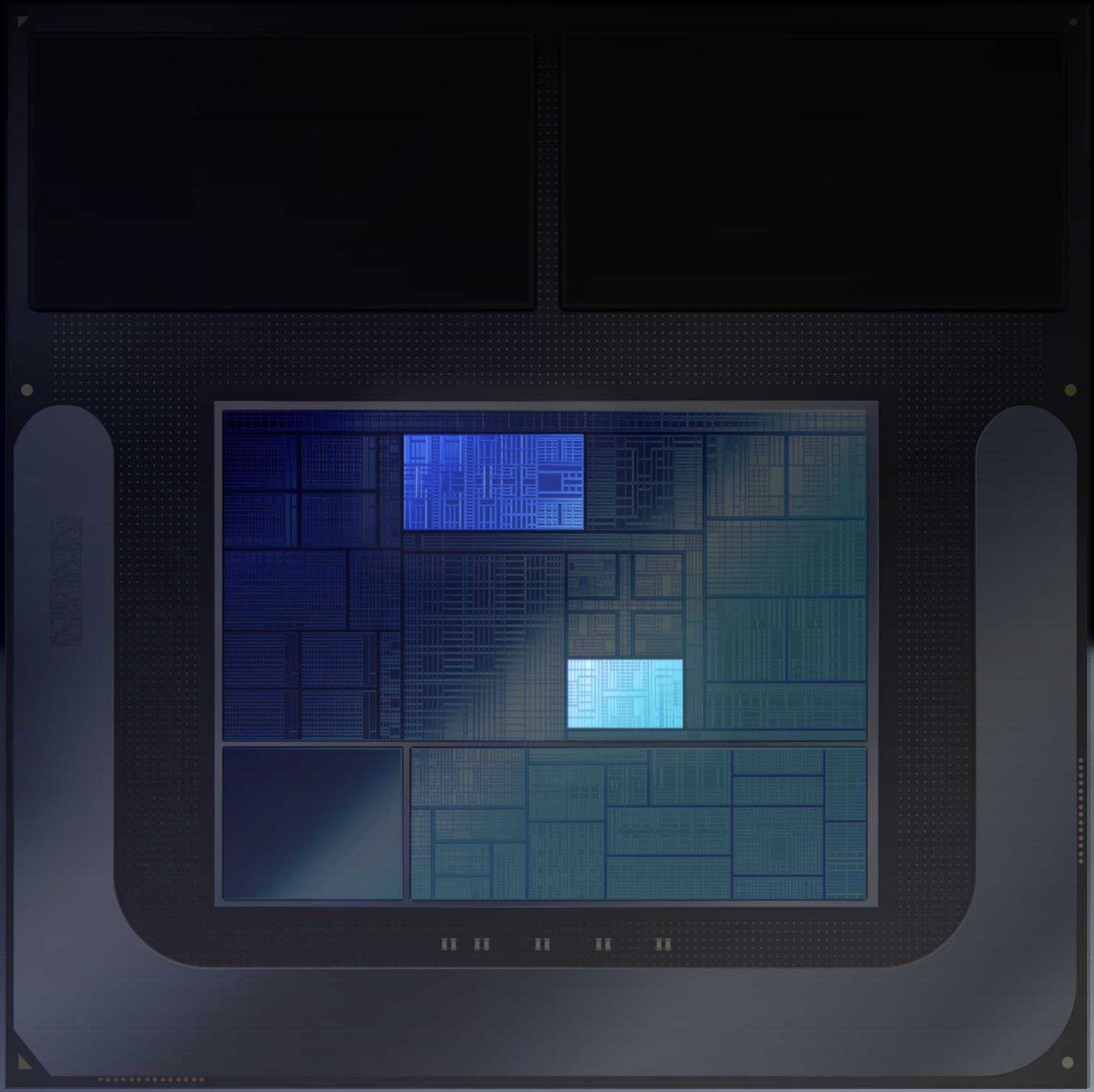
```
C:\Windows\System32>call "C:\openvino\setupvars.bat"
Warning: Python is not installed. Please install one of Python 3.8 - 3.12 (64-bit) from https://www.pyt
downloads/
[setupvars.bat] OpenVINO environment initialized
Press any key to continue . . .
OpenVINO version: OpenVINO Runtime
    Version : 2024.3.0
    Build   : 2024.3.0-15502-66093834e38

Loading and compiling text encoder: 2459.51 ms
Loading and compiling UNet: 8908.48 ms
Loading and compiling VAE decoder:
```

"Older Intel Engineer wearing a blue hat with a blue jacket,
he is making a funny face with his mouth open, excited for new technology."

Lunar Lake

Media & display engines



New Media & Display Engines

Media engine

AV1

Encode & decode

VVC

Decode

Display engine

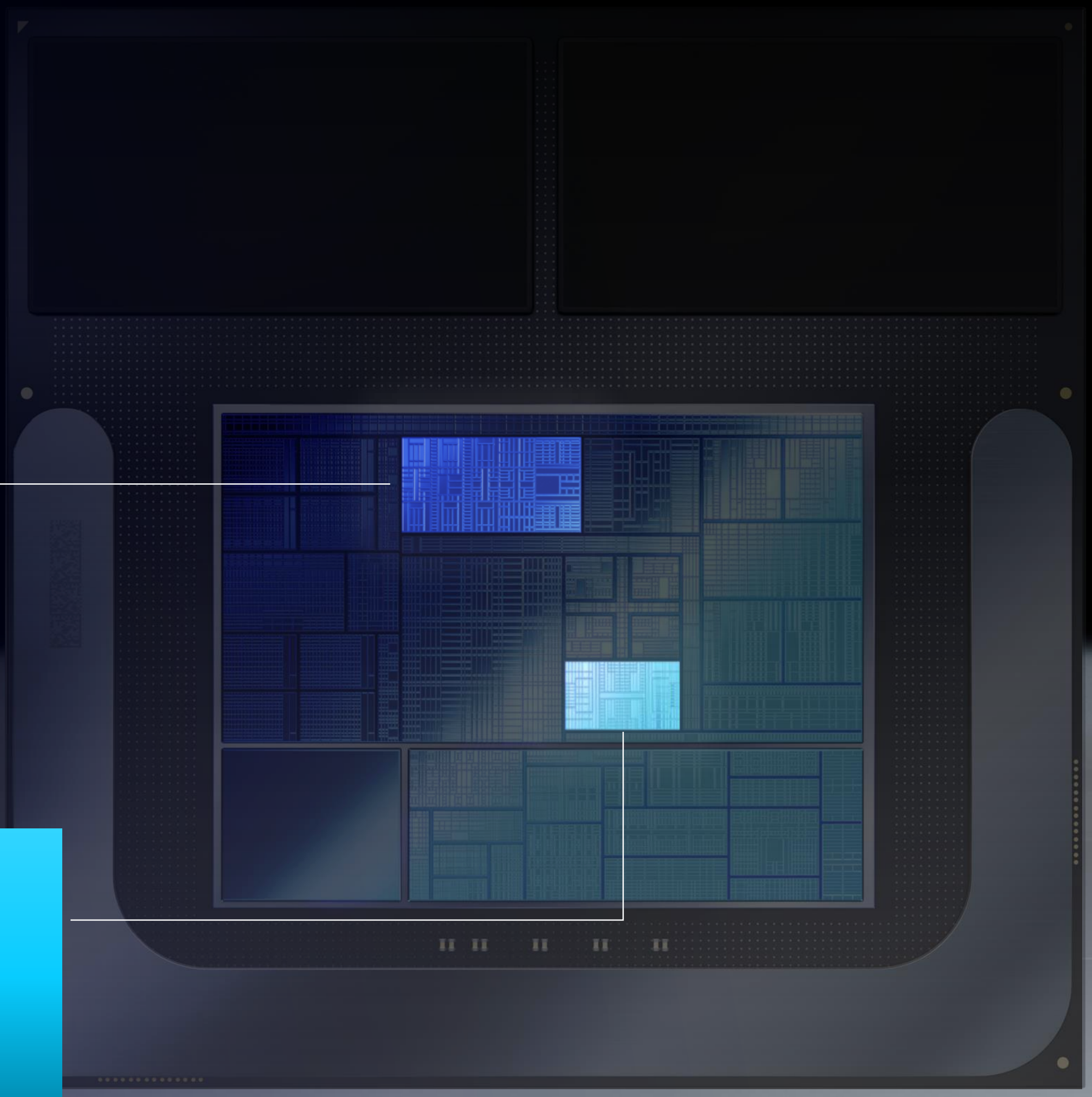
1x eDP 1.5

DisplayPort 2.1

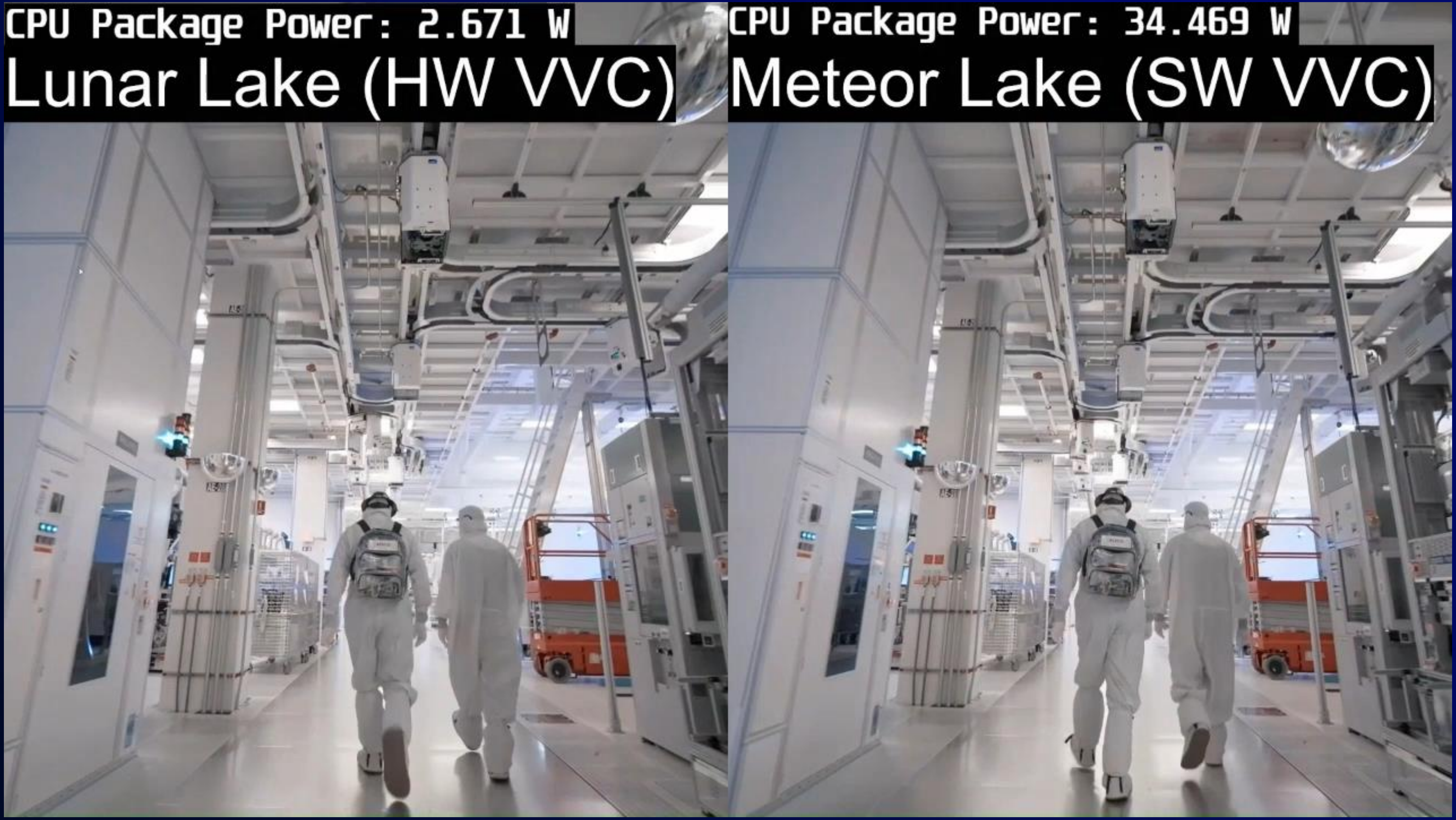
HDMI™ HIGH-DEFINITION MULTIMEDIA INTERFACE

HDMI 2.1

3 display pipes

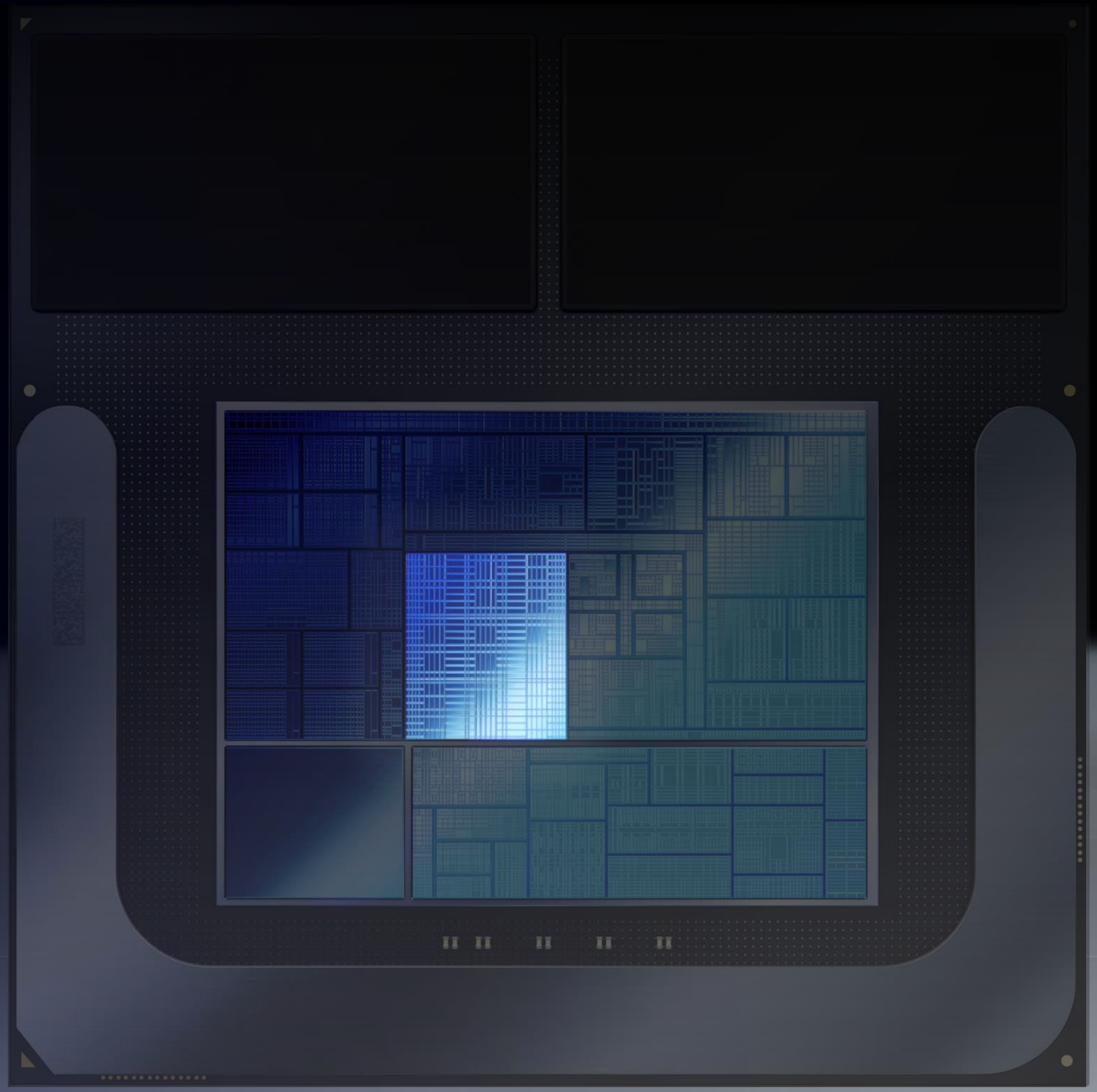


VVC Power reduction



Lunar Lake

New NPU 4.0



Next Gen NPU 4

Architecture goals

Increase NPU size

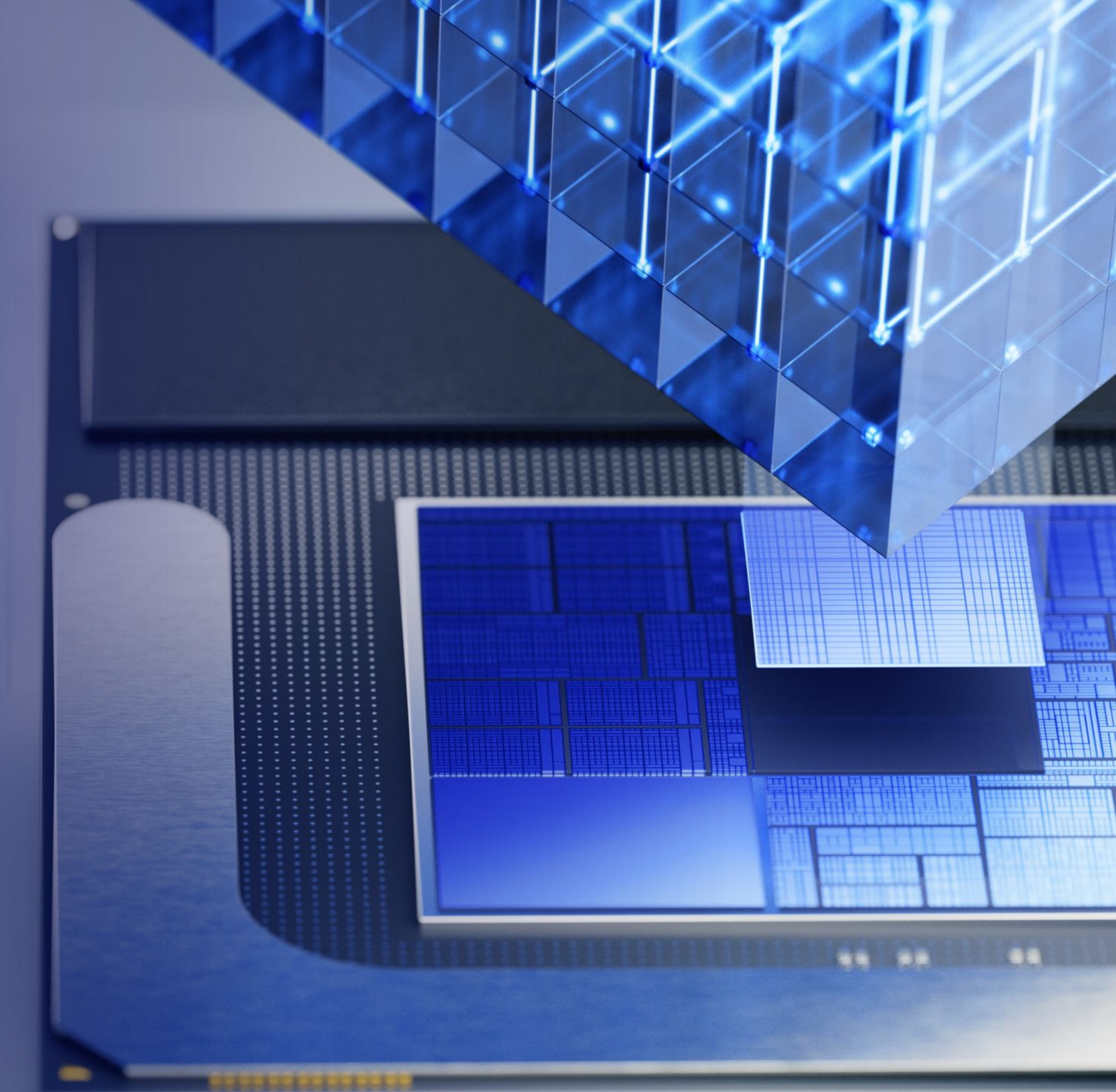
To run next gen AI workloads

Increase clock & efficiency

To increase performance and battery life

Optimize for modern AI

For efficiently running LLMs and transformers

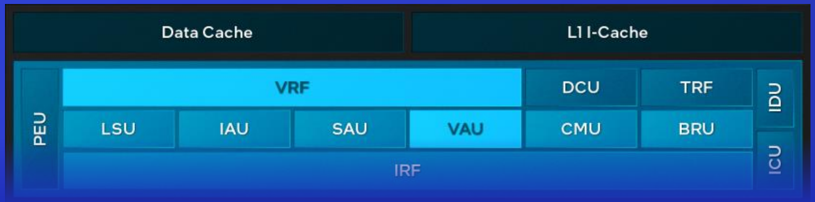


Next Gen NPU 4

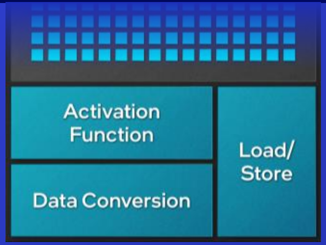
Largest integrated and dedicated AI accelerator for the AI PC

12 **Enhanced SHAVE DSPs**

Accelerating LLM & transformer operations



Native activation function & data conversion support



up to

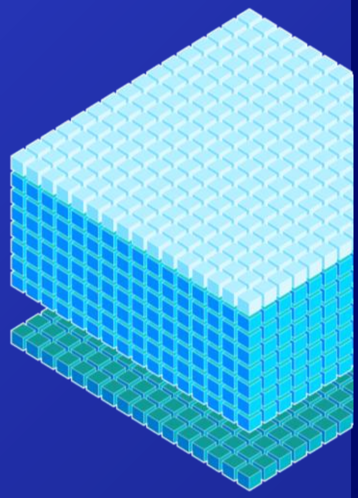
48

TOPS

2x

Bandwidth

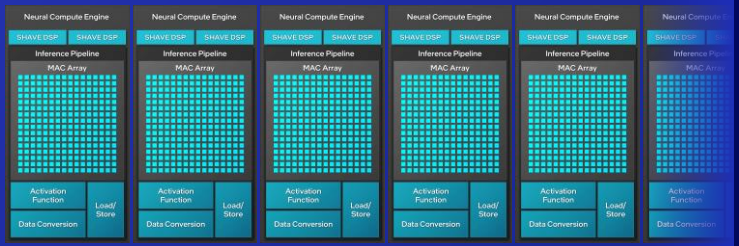
Efficiency optimized MAC array



DMA

Embedding tokenization used for LLMs

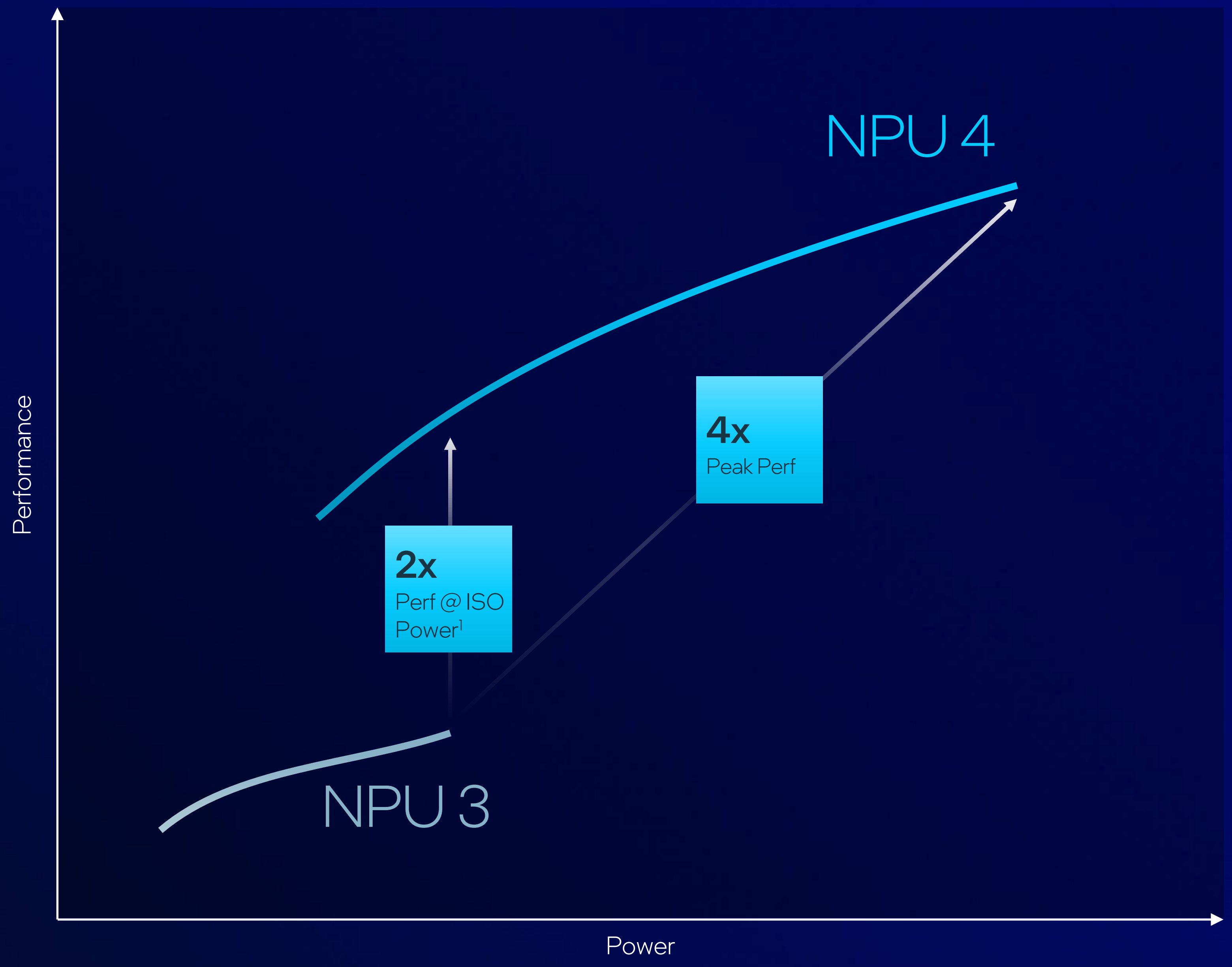
6 Neural compute engines



intel.

Next Gen
NPU 4

Scaling AI performance and efficiency at AI pace



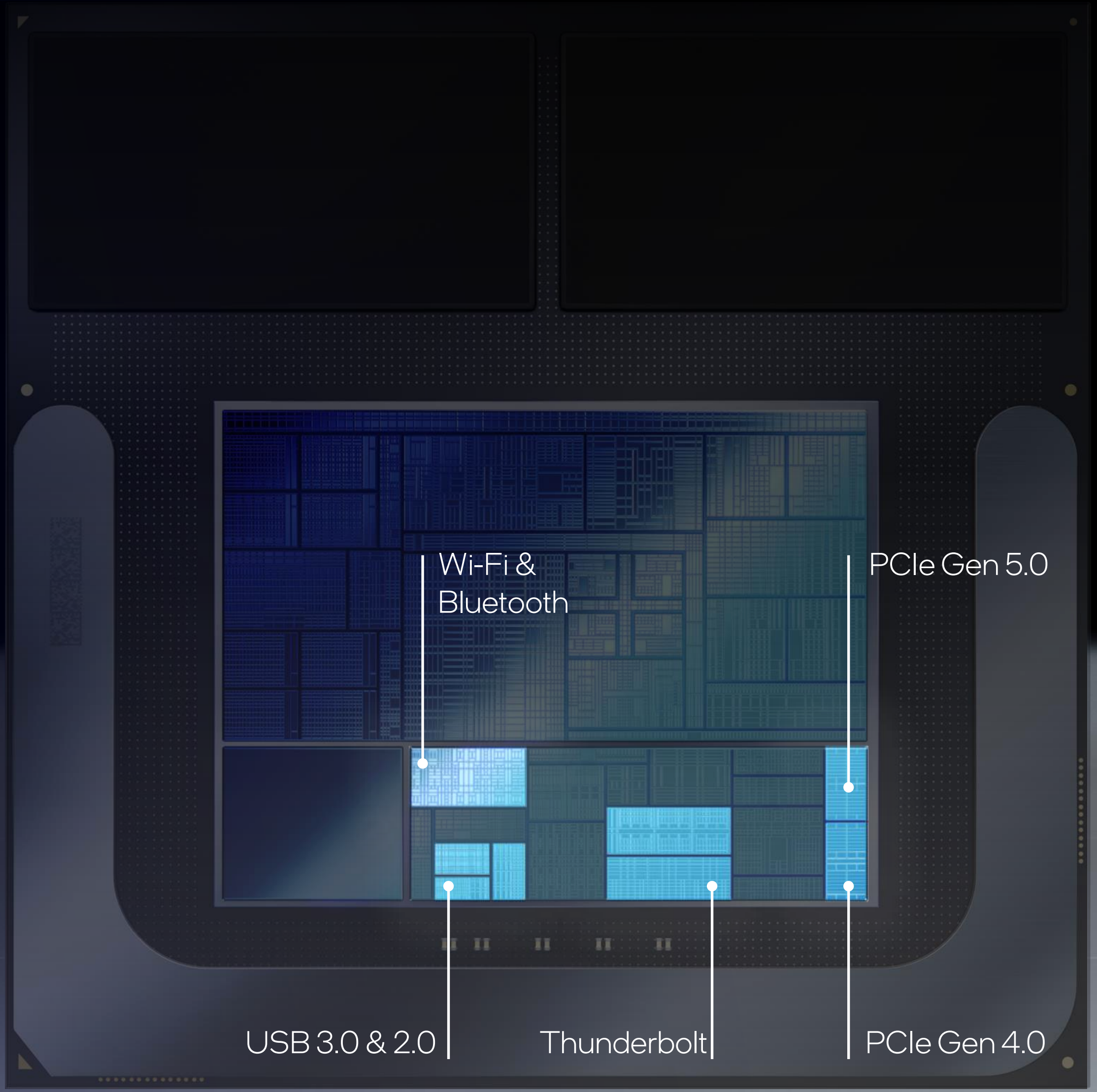
🟦 NPU 4 – in Lunar Lake

⬜ NPU 3 – in Meteor Lake

[1] Based on pre-production simulation data of a real network. See backup for details.

Lunar Lake

Connectivity



Leadership Connectivity

Integrated right onto the package

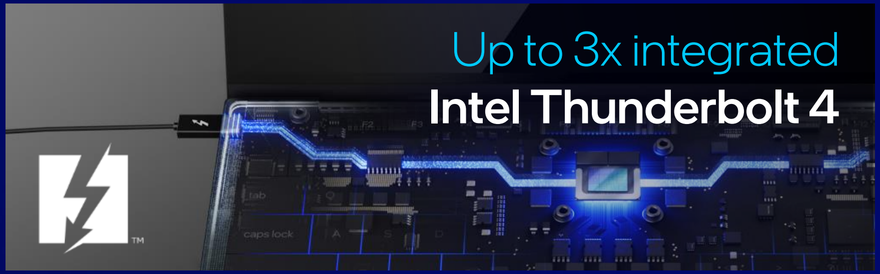
Intel Unison

New multi-device experiences

Tablet control

Swift connect

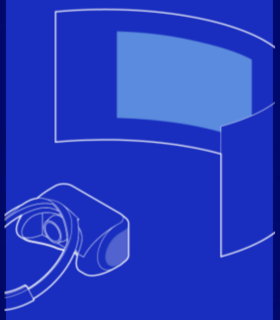
Universal hotspot*



Up to 3x integrated Intel Thunderbolt 4

Thunderbolt Share

Share between PCs at Thunderbolt speed



Enhanced VR

With Wi-Fi 7 & Intel Killer Wi-Fi

Intel® Bluetooth® 5.4

For efficient & HD audio

Integrated Intel® Wi-Fi 7 (5 Gig)

5.8Gb/s

Wi-Fi 7 speed

40Gb/s

TBT4 speed

Up to
4x PCIe Gen 5.0
4x PCIe Gen 4.0

intel.

* Available post LNL launch

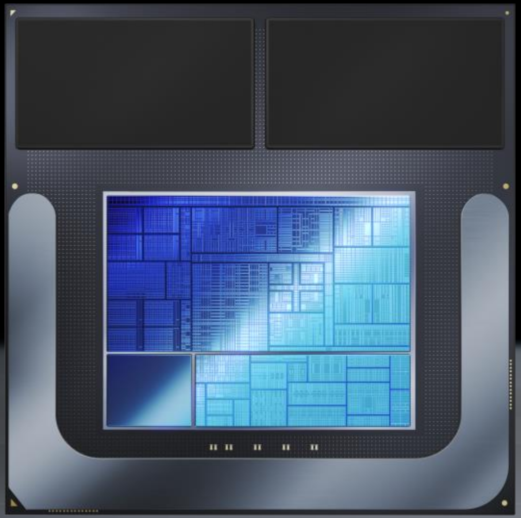
Lunar Lake

Goals

Multi-threaded

Single-threaded

AI compute



Graphics performance

Core perf/watt

Energy efficiency

- >20% MT perf with Lower Core Count
- > 20% Single Thread performance , but more important , delivers the same MTL Single thread performance at half the power.
- More than twice the performance per watt in productivity
- AI throughput of over 3 times the throughput across the NPU, GPU, and CPU
- For graphics apps, Lunar Lake delivers up to 50% uplift in performance, much better user experiences for Gamers and creators
- For battery life, Lunar lake reduces SoC power by up to 40% , major step for mobile in which users will notice

Thank You

Notices & Disclaimers

The preceding presentation contains product features that are currently under development. Information shown through the presentation is based on current expectations and subject to change without notice.

Results that are based on pre-production systems and components as well as results that have been estimated or simulated using an Intel Reference Platform (an internal example new system), internal Intel analysis or architecture simulation or modeling are provided to you for informational purposes only. Results may vary based on future changes to any systems, components, specifications or configurations.

Performance varies by use, configuration and other factors. Learn more at www.intel.com/PerformanceIndex.

AI features may require software purchase, subscription or enablement by a software or platform provider, or may have specific configuration or compatibility requirements. Details at www.intel.com/AIPC.

No product or component can be absolutely secure. Intel technologies may require enabled hardware, software or service activation.

All product plans and roadmaps are subject to change without notice.

Performance hybrid architecture combines two core microarchitectures, Performance-cores (P-cores) and Efficient-cores (E-cores), on a single processor die first introduced on 12th Gen Intel® Core™ processors. Select 12th Gen and newer Intel® Core™ processors do not have performance hybrid architecture, only P-cores or E-cores, and may have the same cache size. See ark.intel.com for SKU details, including cache size and core frequency.

Built-in Intel® Arc™ GPU only available on select Intel® Core™ Ultra processor-powered systems; OEM enablement required.

Some images may have been altered or simulated and are for illustrative purposes only.

Built into the hardware, Intel® Thread Director is provided only in performance hybrid architecture configurations of 12th Gen or newer Intel® Core™ processors; OS enablement is required. Available features and functionality vary by OS.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

While Wi-Fi 7 is backward compatible with previous generations, new Wi-Fi 7 features require PCs configured with Intel Wi-Fi 7 solutions, PC OEM enabling, operating system support, and use with appropriate Wi-Fi 7 routers/APs/gateways. 6 GHz Wi-Fi 7 may not be available in all regions. Performance varies by use, configuration, and other factors. For details on performance claims, learn more at www.Intel.com/performance-wireless.

APPENDIX

Claim # & Statement	Slide # & Title/Details
	SLIDE 2: Flagship SoC for the next gen of AI PCs
Up to 40% lower SoC power	Testing by Intel as of May 2024. Data based on Lunar Lake reference validation platform measurement vs Meteor Lake reference validation platform as measured by YouTube 4K 30 AV1.
Similar ST perf at half the power	Testing by Intel as of May 2024. Data based on Lunar Lake reference validation platform as measured CBR24 ST vs. prior generation.
Up to 1.5X better graphics	Testing by Intel as of May 2024. Data based on Lunar Lake reference validation platform measurement vs Meteor Lake reference validation platform as measured by 3DM Time Spy. 3DMark*
	SLIDE 9: Lion Cove P-core
Lion Cove core delivers 14% better IPC vs. Redwood Cove core	Iso frequency benefit estimate across: components of SPECrate2017_int_base and SPECrate2017_fp_base (estimated) running 1 copy
Lion Cove Performance at different power levels vs. Redwood Cove	Data based on measurements on an Intel internal reference validation platform. Performance: SPECrate2017_int_base (estimated) running 1 copy. Power: Fixed PL1

APPENDIX

Claim # & Statement	Slide # & Title/Details
	SLIDE 13: Skymont E-core
Skymont IPC on Lunar Lake Low Power Island: 1.38x integer and 1.68x floating point vs. Meteor Lake LP E-core (Crestmont)	Results are based on Intel's internal projections/estimates as of 5.13.2024(+/- 10% Margin of Error) on SPEC CPU 2017 Rate est, GCC12.1-O2 Linux at Fixed Frequency (ISO).
Skymont Power & Performance on Lunar Lake Low Power Island: up to 2x peak ST performance or 1/3 the power at similar ST performance and up to 4x peak MT performance or 1/3 of the power at similar MT performance	Results are based on Intel's internal projections/estimates as of 5.13.2024(+/- 10% Margin of Error) on SPEC CPU 2017_int_base est, GCC12.1-O2 Linux (ISO). Comparing a Skymont E-core cluster (4 Skymont cores) vs. Meteor Lake LP E-core cluster (2 Crestmont cores) to showcase workload coverage increase for the Lunar Lake Low Power Island.
	SLIDE 15: Breakthrough x86 Power Efficiency
>50% peak performance >20-80% per/watt	Illustration of the relative Lunar Lake P-core and E-core performance across the SoC power range.
	SLIDE 21: Improved Experience
35% power reduction when containment & power management optimization are enabled	As of May 2024, based performance estimated with measurements on Lunar Lake reference platform with power optimizations enabled vs. power optimizations disabled.

APPENDIX

Claim # & Statement	Slide # & Title/Details
	SLIDE 26: Next Gen Xe2 GPU
1.5x better vs. Meteor Lake GPU	Testing by Intel as of May2024. Data based on Lunar Lake reference validation platform measurement vs Meteor Lake reference validation platform as measured by 3DM Time Spy. 3DMark*
	SLIDE 28: Next Gen Xe2 GPU
1.5x graphics performance over Meteor Lake	Testing by Intel as of May2024. Data based on Lunar Lake reference validation platform measurement vs Meteor Lake reference validation platform as measured by 3DM Time Spy. 3DMark*
	SLIDE 37: Next Gen NPU 4
2x performance at ISO power vs. Meteor Lake	Testing by Intel as of January 2024. Based on VPU-EM simulation. Power data is generated from the simulation tool based on power data that has been extracted from circuit simulation tools. This simulation, which is a ~100% utilization int8 network, is expected to correlate well with silicon.
4x peak performance	4x peak performance is based on TOPS increase from MTL (11 TOPS) to LNL (48 TOPS).
	SLIDE 41: Energy Efficiency
Up to 40% lower SoC power vs. Meteor Lake	Testing by Intel as of May 2024. Data based on Lunar Lake reference validation platform measurement vs Meteor Lake reference validation platform as measured by YouTube 4K 30 AV1.

intel®