Inside the 6th Gen Intel® Skylake Core –

Past, Present, and Future of a new microarchitecture

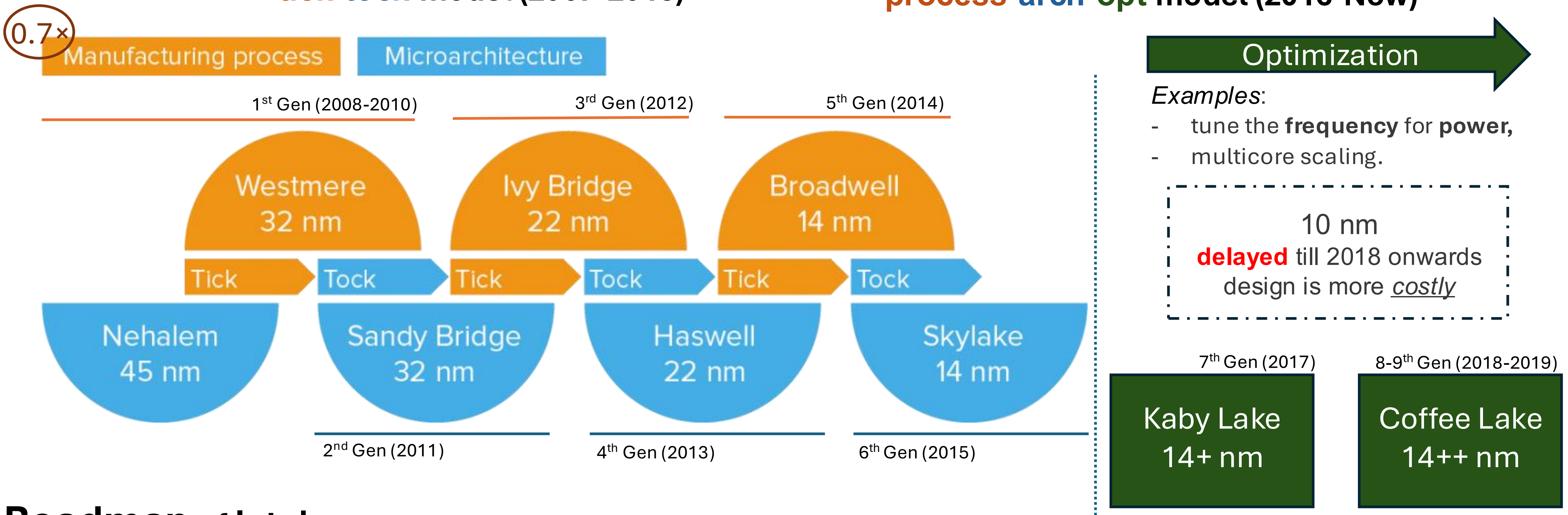
Peter Hu (zh369)

Magdalene College

@ Advanced Topics in Computer Architecture, 3rd Feb 2026

tick-tock model (2007-2015)

process-arch-opt model (2016-Now)



0.7×

Manufacturing process

Microarchitecture

1st Gen (2008-2010)

3rd Gen (2012)

5th Gen (2014)

Westmere 32 nm

Ivy Bridge 22 nm

Broadwell 14 nm

Tick → Tock → Tick → Tock → Tick → Tock

Nehalem 45 nm

Sandy Bridge 32 nm

Haswell 22 nm

Skylake 14 nm

2nd Gen (2011)

4th Gen (2013)

6th Gen (2015)

Optimization

Examples:

- tune the **frequency** for **power,**
- multicore scaling.

10 nm
delayed till 2018 onwards
design is more *costly*

7th Gen (2017)

8-9th Gen (2018-2019)

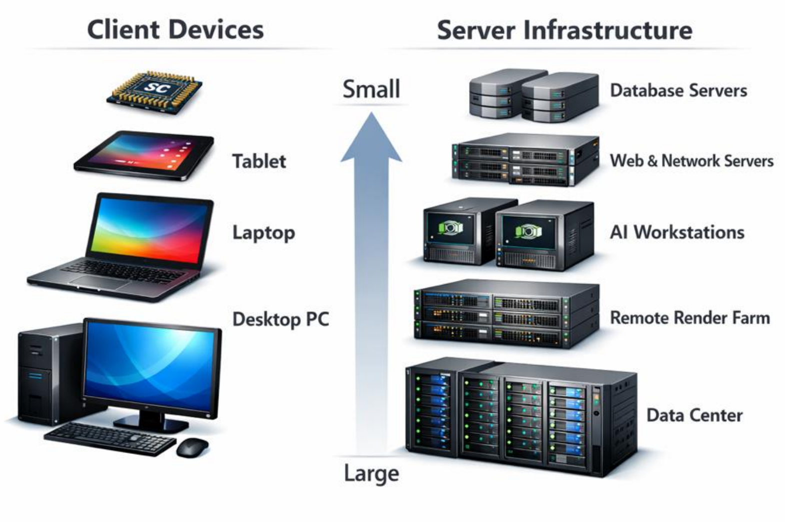
Kaby Lake 14+ nm

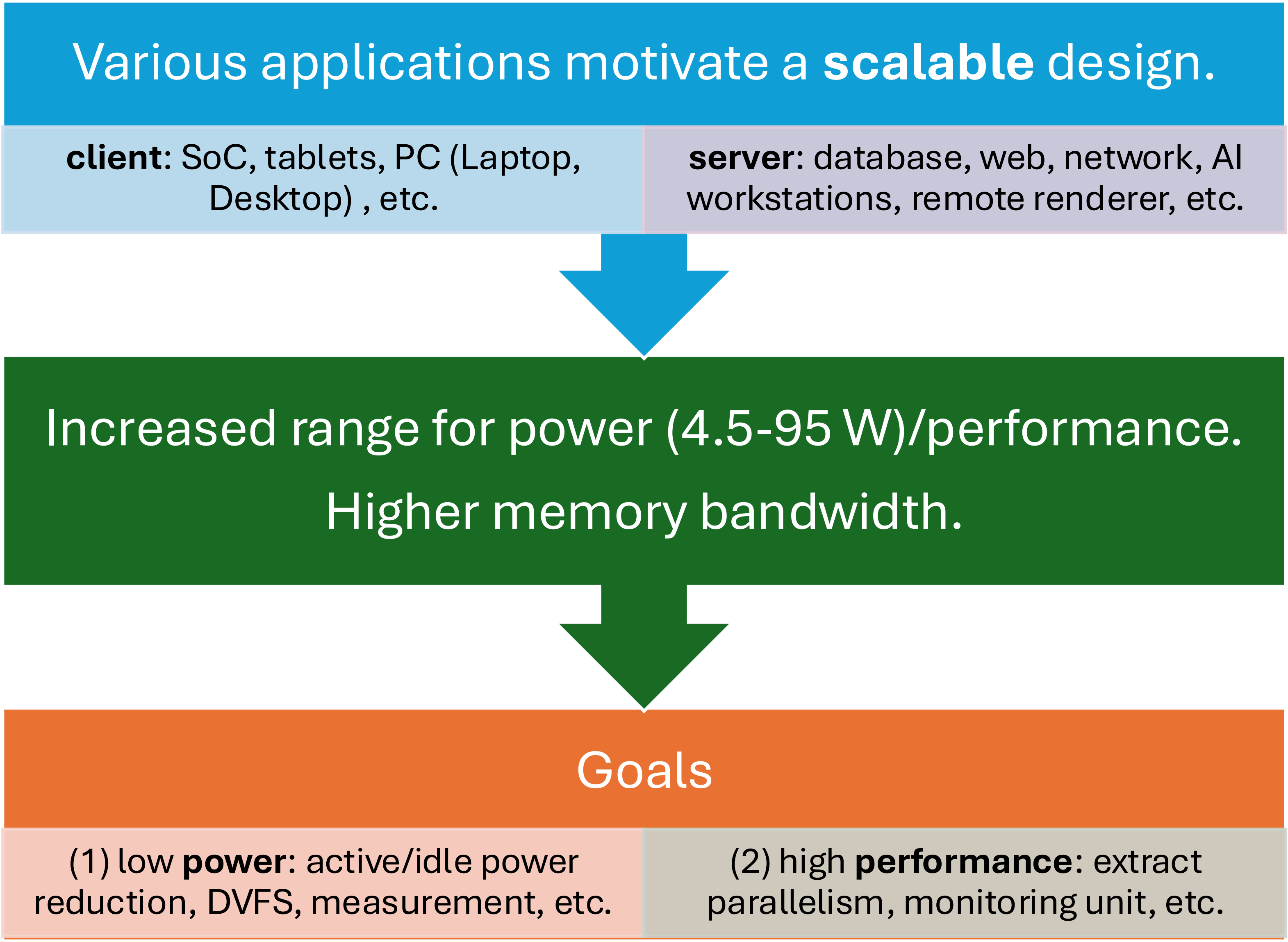
Coffee Lake 14++ nm

Roadmap of Intel *product release*

https://en.wikipedia.org/wiki/Tick%E2%80%93tock_model

Requirements





Various applications motivate a **scalable** design.

client: SoC, tablets, PC (Laptop, Desktop) , etc.

server: database, web, network, AI workstations, remote renderer, etc.

Increased range for power (4.5-95 W)/performance.

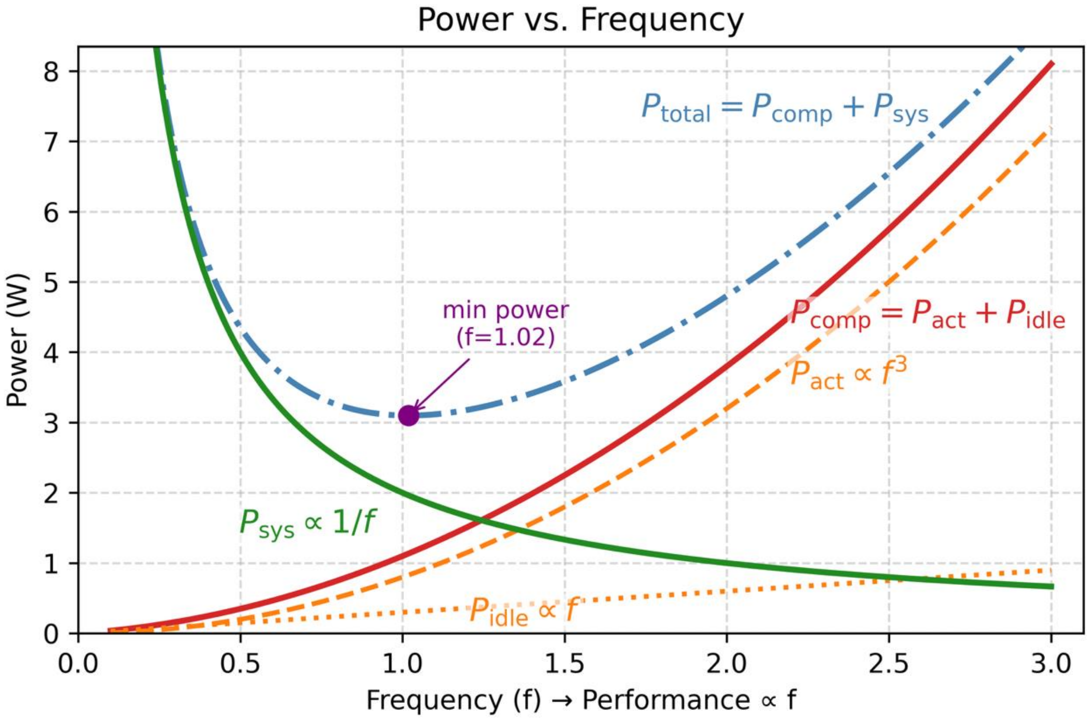
Higher memory bandwidth.

Goals

(1) low **power**: active/idle power reduction, DVFS, measurement, etc.

(2) high **performance**: extract parallelism, monitoring unit, etc.

Power management

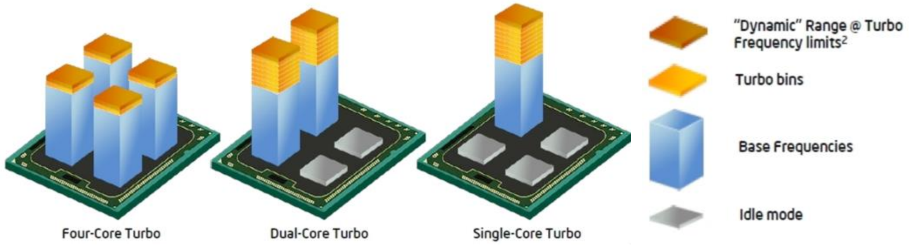


most energy-efficient **balanced** **highest performance**

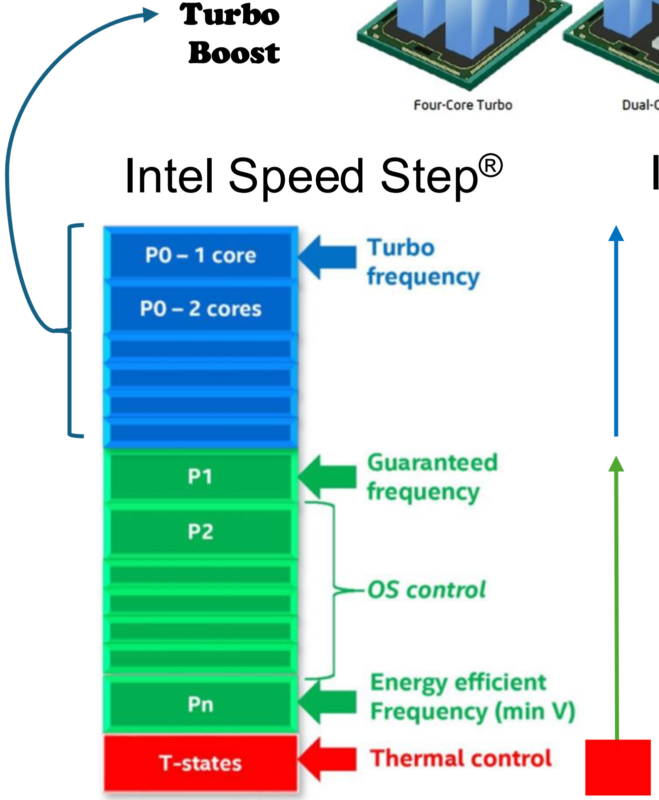
P^{α}

calculated runtime by *Energy Aware* **Race to Halt**

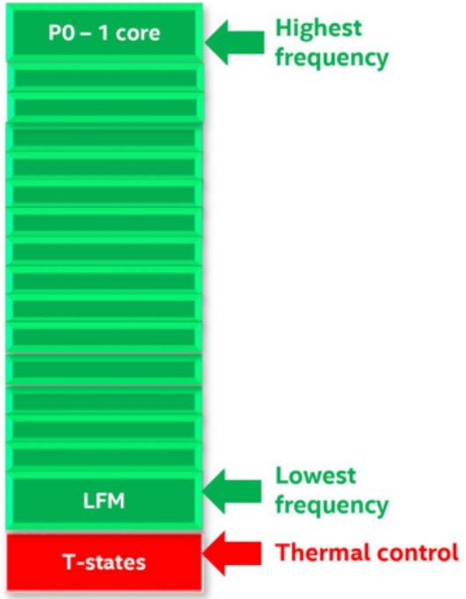
Multi-objective **optimization** $\min_{f} Energy(f) \times Delay^{\alpha}(f)$

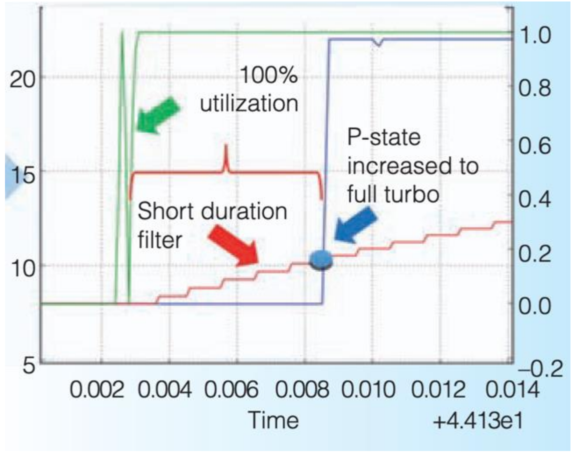


Intel Speed Step®



Intel® Speed **Shift** (**Skylake**)





additional support for

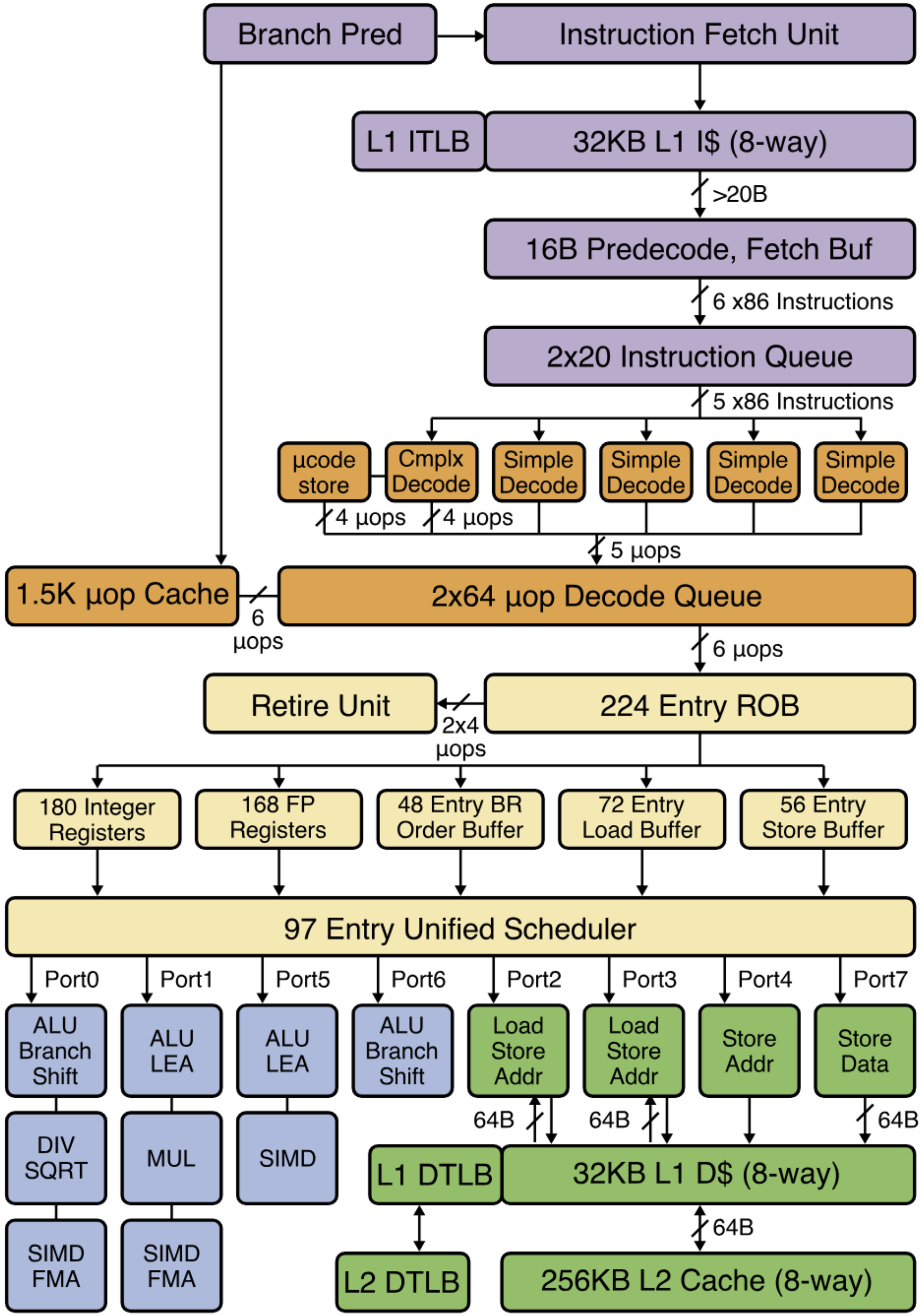
- waiting consumer threads
- responsiveness HW-DVFS
  - 14% improvements

	SpeedStep®	**Speed Shift**
Managed by	software	hardware
P-state from	OS (ACPI governor)	CPU hardware
Latency	ms-scale	**µs-scale**
Responsiveness	moderate	**high**
Runtime behaviour and Microarch. state (memory)	limited visibility	**more information available**
Power efficiency	governor-dependent	**better**

Future: (1) more Boost heuristics (Panther Lake), (2) useful hints from OS, (3) workload prediction with AI.

4TH Gen Micro Arch.
Skylake

Feature (unit)	Nehalem	Sandy Bridge	Haswell	Skylake
BR predictor	BTB **+** two-level predictor for history: 32B Global Buffer + Pattern Table			
Decoders (/cycle)	4	4	4	**5**
Queue (per thread)	28	28	56 total	**64**
Reorder buffer (ops)	128	168	192	**224**
Integer/FP Rename	In ROB	160 / 144	168 / 168	**180 / 168**
ld/st buffer (entries)	48 / 32	64 / 36	**72** / 42	**72 / 56**
Scheduler (entries)	36	54	60	**97**
Issue width	4	6	**8**	**8**
Arithmetic units	throughput increases, latency deduces			
Load	**reduce** store-to-load forward, split-load cost			
Store	deep buffers, request to L2 for earlier L1 miss			
Cache	bandwidth **move** from L2 to shared L3			



Pros & Cons

- overall enhancements over Intel legacy **superscalar** pipeline.
  - **wider frontend** + **deeper backend** and **issue queue**.
- **out-of-issue** provides **higher MLP, ILP.**
- **Better arithmetic.**
  - **SIMD** exploits the **DLP**.
- **Improved memory subsystem**.
  - Larger L3 (LLC), eDRAM, etc.

- diminishing returns due to ILP limits, BR mispredict, memory stalls.
- power increases, to be tuned by 7th Gen (Kaby Lake).
- **speculative exec.** enables spectre attack.
  - solution: clear related cache when speculation failed to **tradeoff** security over performance.
- incremental novelty, no fundamental different microarchitecture design.

Fabric, Cache, and Memory

four **rings**: request, snoop, data, acknowledge.
high bw, low latency/power, modular, less control logic.

scalable fabric, shared LLC,
larger bandwidth, more coherency.

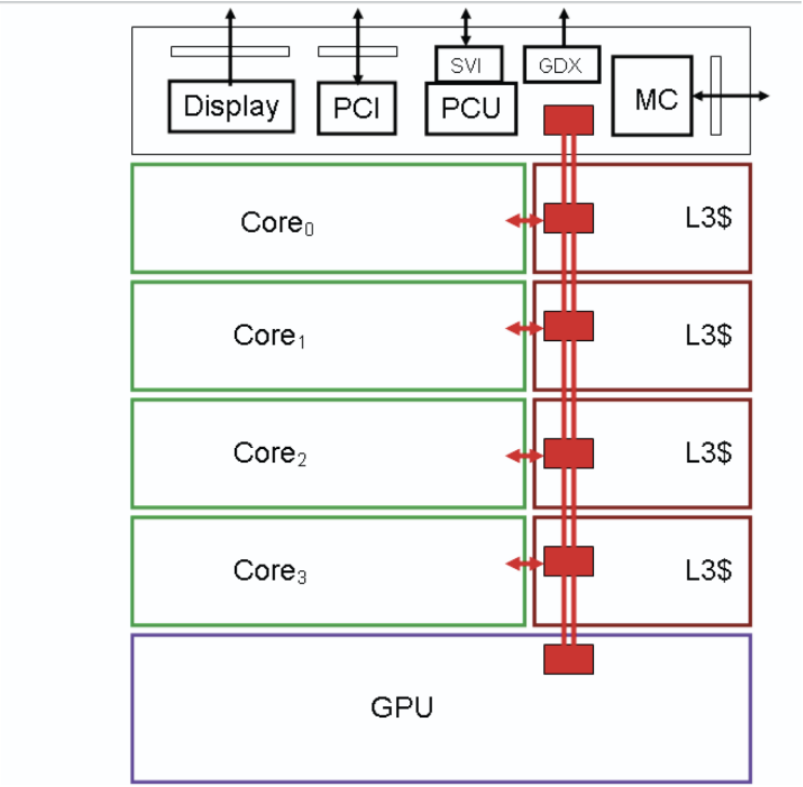
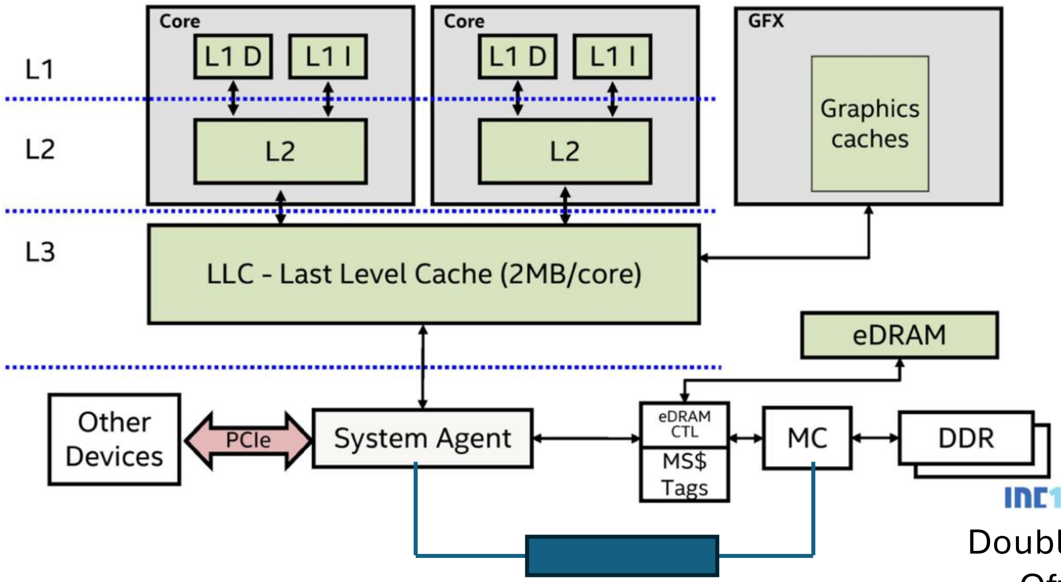


Figure 15.1.1: Sandy Bridge block diagram.

Skylake memory subsystem



DDR3, DDR4, LPDDR3.

Double Data Rate
Off-chip DRAM

Memory Encryption Engine
(*MEE*)

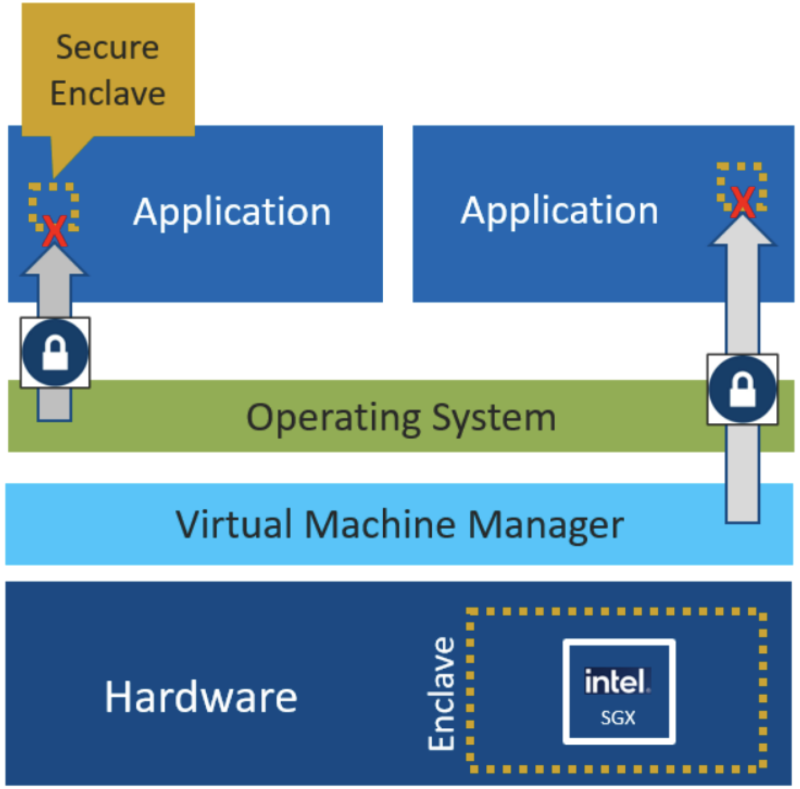
Weakness: poor scalability, contention for multi-core, clocking.
For large dies used in server, **mesh** later from 2017 replaced the ring.

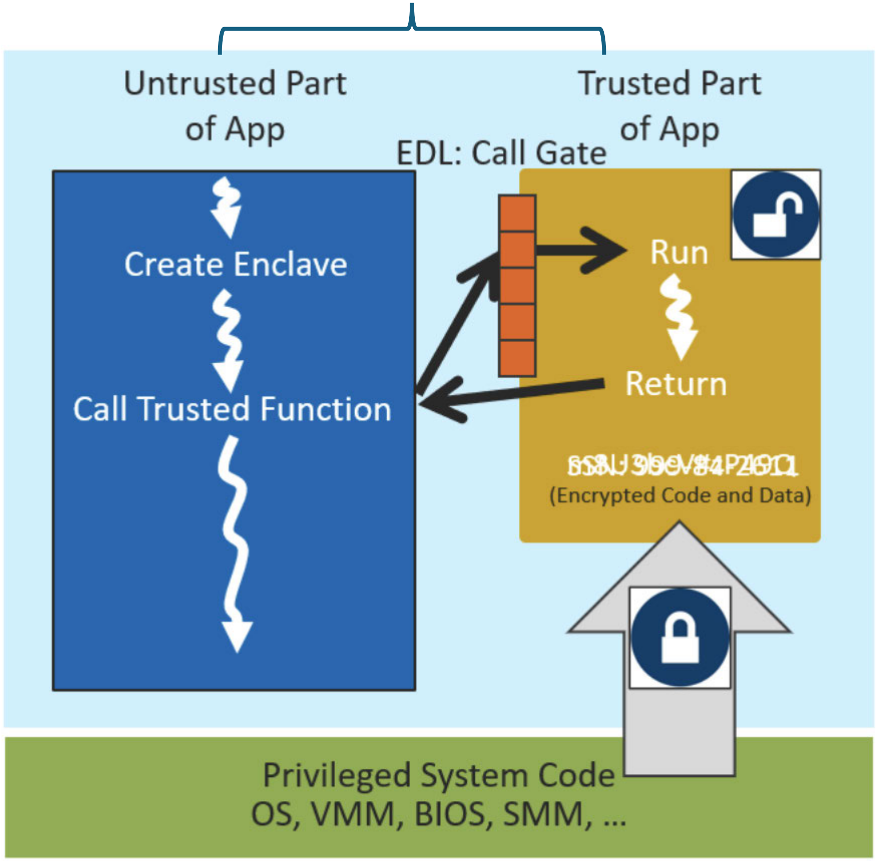
Memory security & integrity

Software Guard Extensions (SGX) : a new ISA extension

protect DRAM against **memory bus snooping** and **cold boot attacks** for enclave code and data.
usage: key management, password vaults, secure analytics, confidential ML inference.

removed from trust boundary

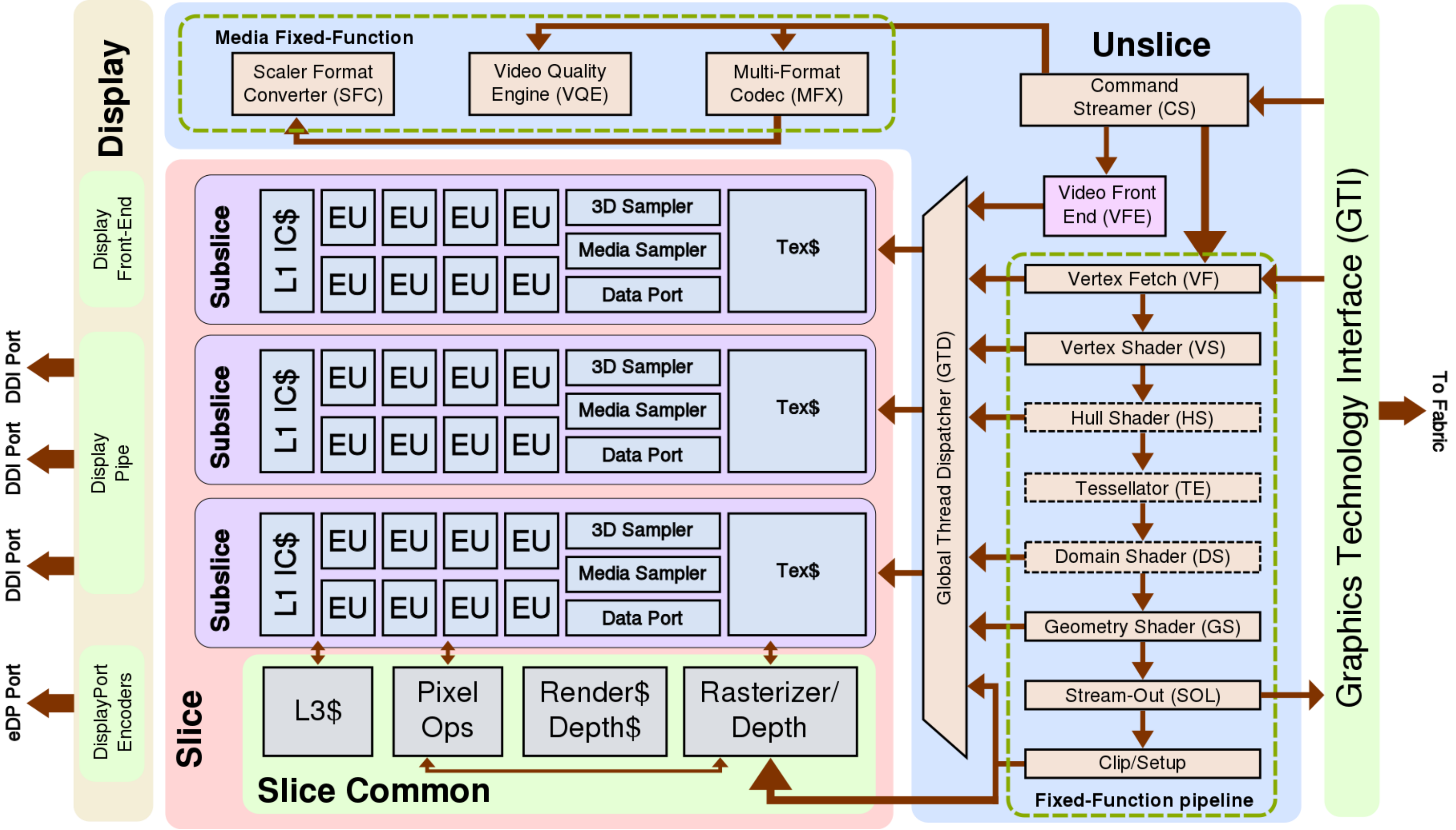


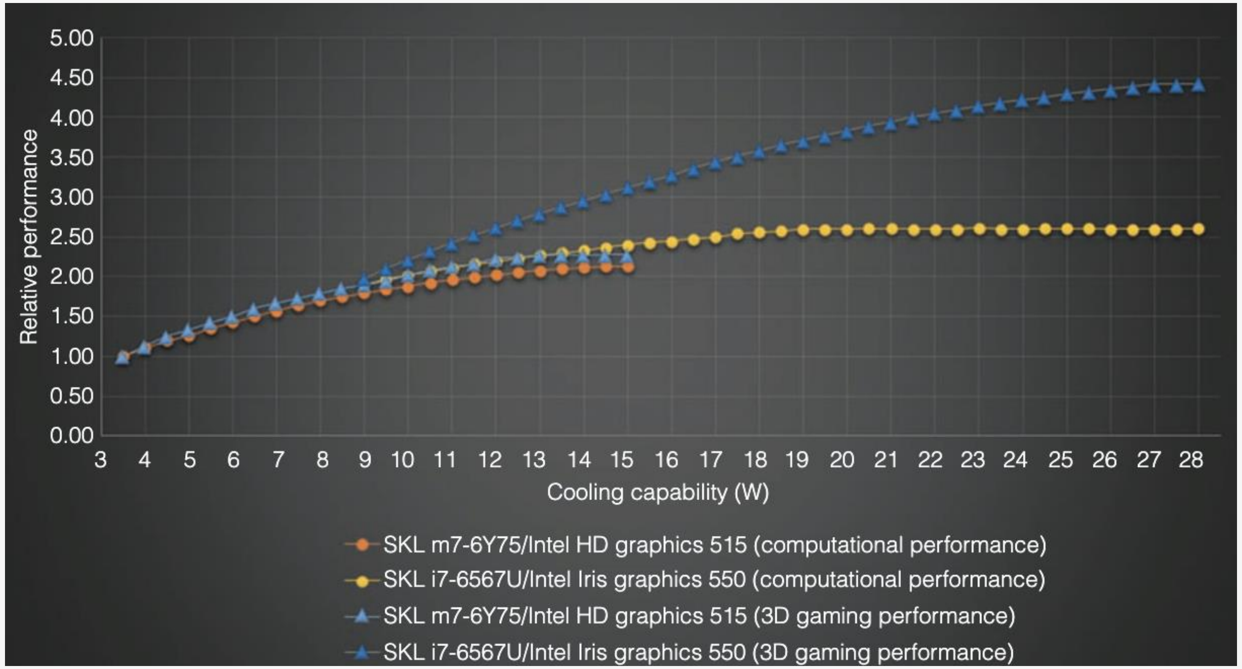


Weakness: limited size, performance overhead (sys call), side-channel attacks (cache).

Graphics

- Lossless render compression,
- Low-power media, video quality,
- Slice vs. Unslice,
  - different clock domains, power-gate.
  - pure media, higher throughput.
- Idle management,
  - Voltage-Frequency (V-F) curves,
  - C-state: C0 (fully active), C1 (halt), C2 (clocks off), etc.



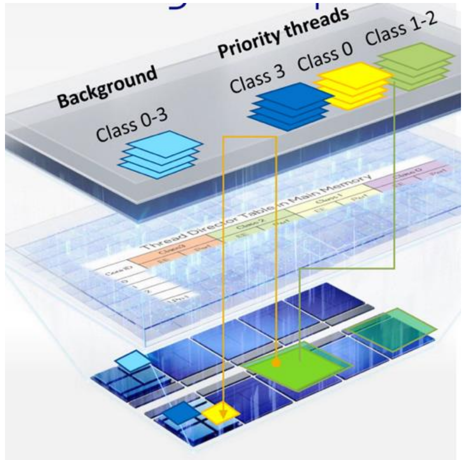


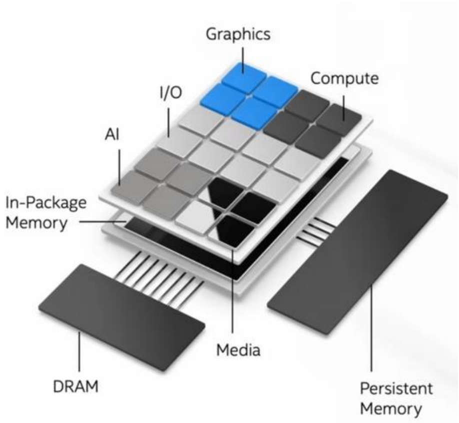
Intel® **longest-serving** CPU architecture: Skylake.

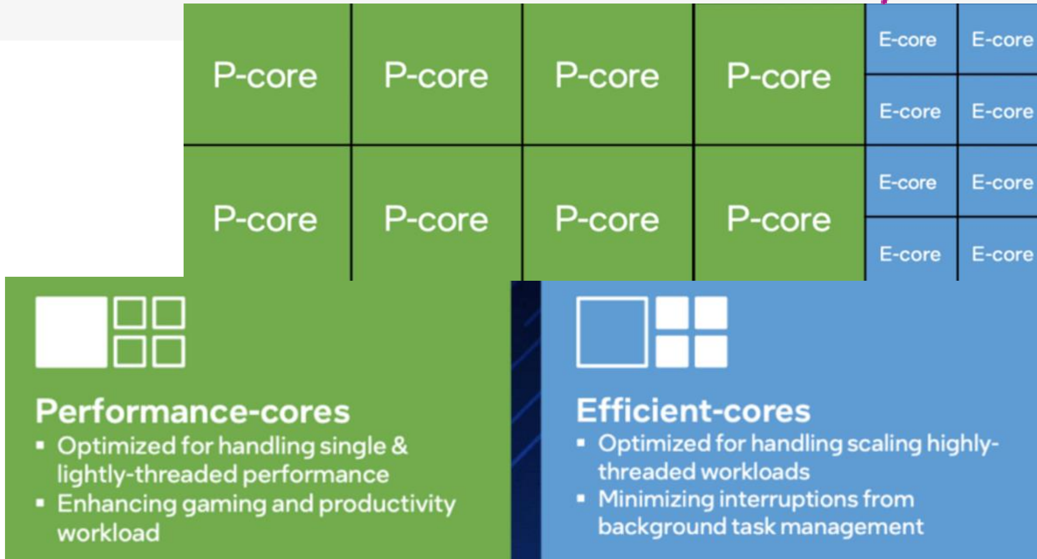
- *Incremental* novelty in
  - power management,
  - superscalar microarchitecture,
  - fabric, interconnect and memory system.
- *Evolutionary* novelty in
  - memory encryption (SGX),
  - decoupling of slice and unslice graphics unit.
- **Results**: meet the thermal/power constraints, with
  - better computational & 3D performance,
  - longer battery life,
  - wider cooling capability.

Post-Skylake changes

Efficiency-first design choices	shallower pipelines, **reduce freq.** for energy (10th Ice Lake / 12th Alder Lake).
Front-end width	wider fetch/decode, higher µOP delivery rate to reduce bubbles.
Instruction fetch	larger BTBs, deeper history; smarter prefetching.
Instruction window	ROB and scheduler size increased → more MLP & ILP.
Execution back-end	more ports, AGUs, load/store bandwidth.
Larger private L2 caches	reduced L3 traffic, better core-locality.
Heterogeneous cores	**P+E:** High IPC cores for ***perf.*** + small cores for ***energy*** (12th Alder Lake+).
HW-DVFS (Speed Shift)	quicker frequency changes, aggressive.
HW-OS co-DVFS	**Thread Director** guides workload-aware freq. steering (12th Alder Lake).
Block- & tile-based gating	independent power/thermal control **per block** (Meteor Lake+, 2023).







Performance-cores

- Optimized for handling single & lightly-threaded performance
- Enhancing gaming and productivity workload

Efficient-cores

- Optimized for handling scaling highly-threaded workloads
- Minimizing interruptions from background task management

Thank you and welcome your thoughts!

Reference

- Inside 6th-Generation Intel® Core: New Microarchitecture Code-Named Skylake, Jack D. et al., IEEE Micro, 2017.
- Intel® SGX – Key Management on the 3rd Generation Xeon® Scalable Processor, 2021.
- Gen9 - Microarchitectures – Intel®, Wiki-Chip.
- Skylake case study, Prof Robert Mullins, Cambridge Uni Advanced Computer Arch.
- Computer arch Intel® Skylake, Prof Christopher Batten, Cornell University ECE 4750.
- Spectre Attacks: Exploiting Speculative Execution, Paul K. et al., 2019 IEEE Symposium on Security and Privacy (SP).
- A Fully Integrated Multi-C/GPU and Memory Controller 32nm Processor, Marcelo Y. et al., Intel Sandy Bridge, 2011.
- The 1st to latest Gen. Intel® architectural design details, from various documentations, reviews, videos, blogs, etc.