TWENTY-SEVENTH ANNUAL

TestConX™

Archive

DoubleTree by Hilton
Mesa, Arizona
March 1-4, 2026

ISE Labs | ASE

Burn-In at an Inflection Point

Supporting HPC, Chiplets, and CPO in Production Test

Vijay Israni

Keynote TestConX 2026

TestConX™

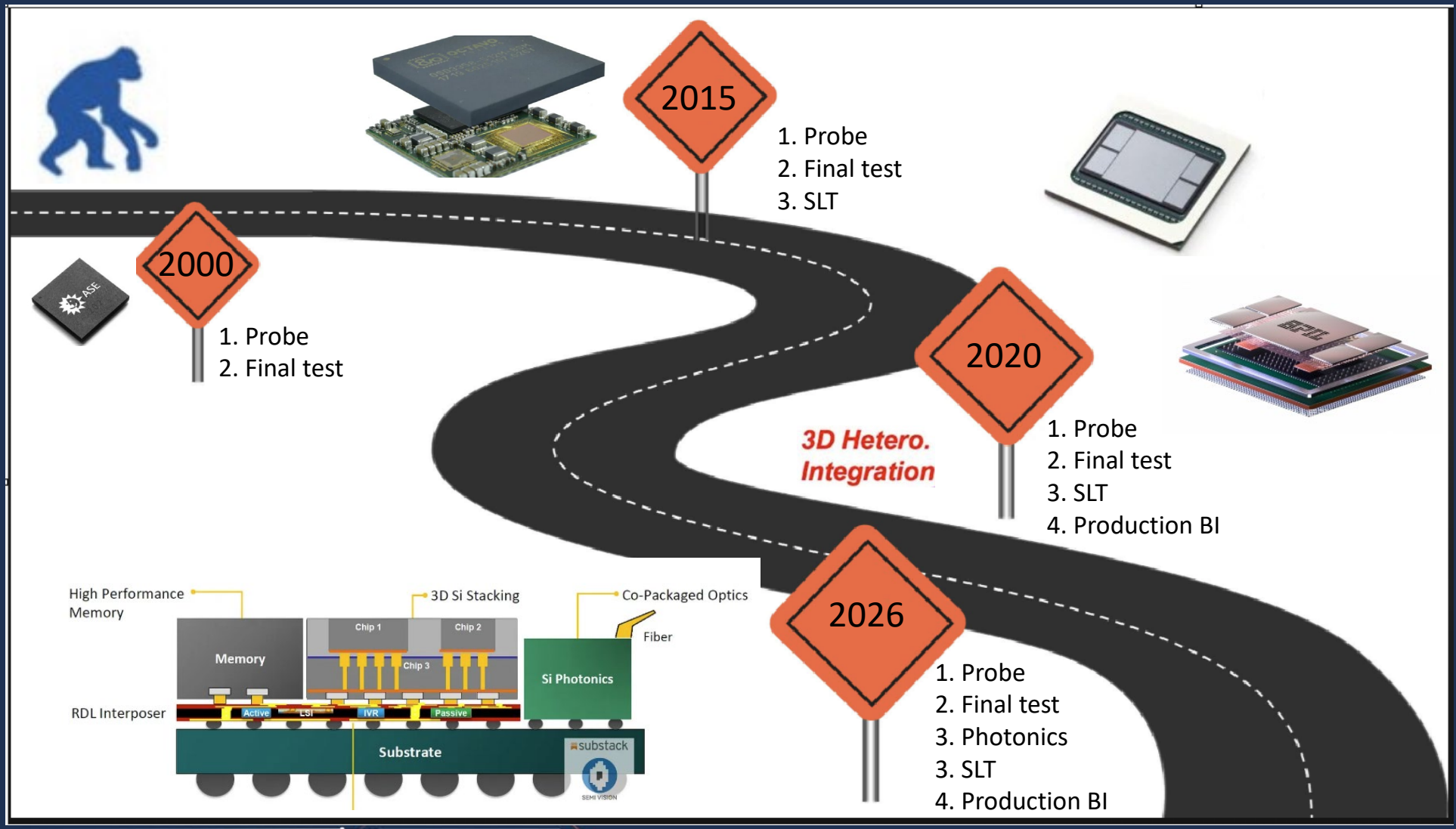
2026.03.01

Who or what is leading the industry to the inflection point?

HPC → AI

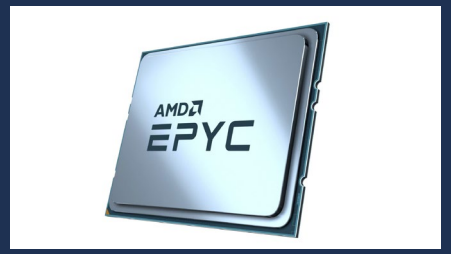
HPC (High Performance Computing) products Powering the AI Boom

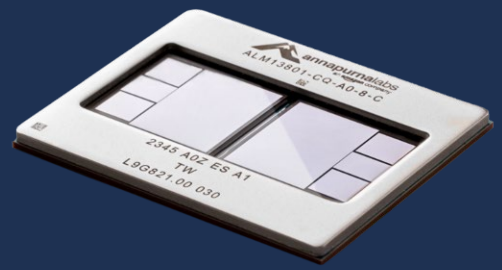
Transformation of Semiconductor Test Criteria for HPC

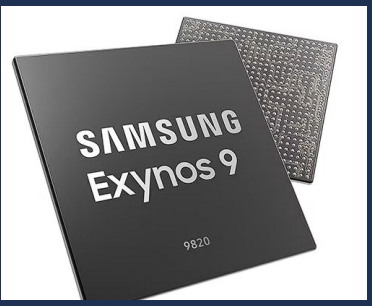


- Wafer probe Initial functional screening at the die level (3x solutions PIC, EIX, & ASIC)
- Final Test Post-packaging test for performance, yield, and binning
- SLT Emulates real-world conditions to catch system-level issues
- Production BI Applies thermal and electrical stress to screen for early-life failures
- Wafer level burn-in finds parameter shifts & weak vias.
- Optical Engine tests confirms Known-Good-Chiplet
- Silicon Photonics Testing introduces optical/electrical co-testing requiring new infrastructure for alignment, signal integrity, and high-speed data validation

Products that make up HPC







GPUs	CPUs	ASICs	NPUs
AI training backbone	System orchestration	Custom AI acceleration	On-device AI (typically embedded)
NVIDIA Blackwell	Intel Xeon Scalable	Google TPU	Apple Neural Engine
AMD Instinct MI300/350	AMD EPYC	AWS Trainium	Qualcomm Snapdragon Elite (Hexagon)
		Cerebras WSE-3	Tensor NPU

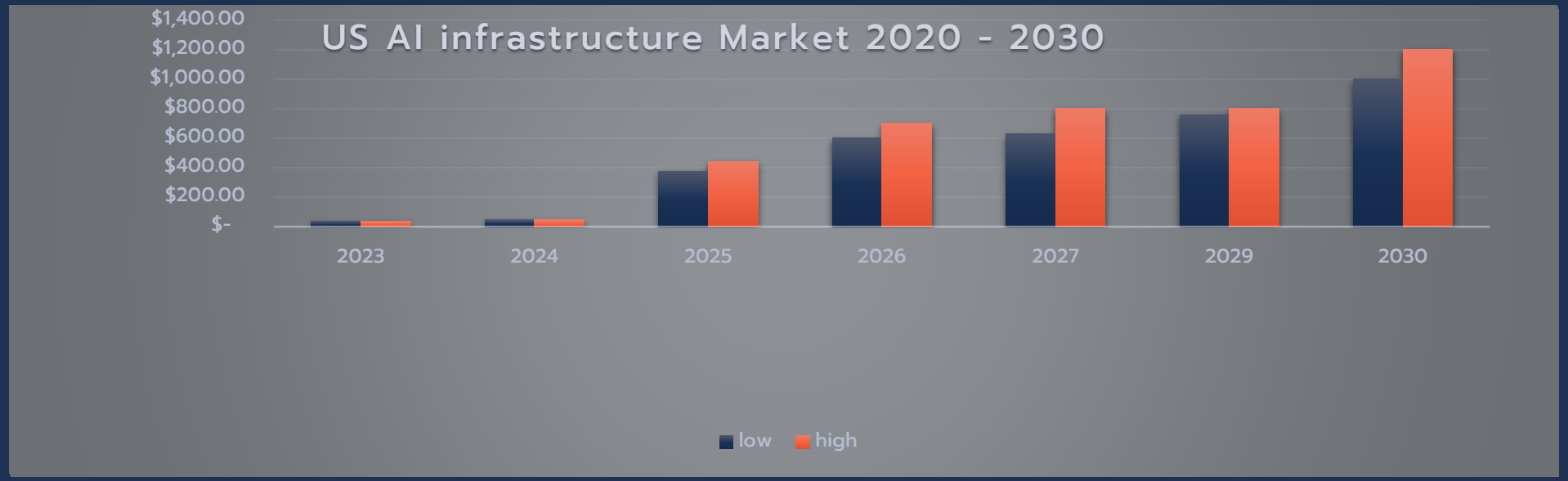
HPC – The Inflection Point broken out by Segment (in Billions)

Segment	Incl	Today (2024/2025)	Outlook (2030)	Source
AI Accelerators	GPUs + AI ASICs/ASSPs for cloud/DC (training & inference)	**$207**	**$286**	Omdia
AI Inference	HW & platforms serving inference workloads DC/cloud/edge	**$160**	**$255**	MarketsandMarkets
AI training	Training portion of DC GPU/accelerator spend	**$74**	**$140**	MarketsandMarkets
Data Center networking	Ethernet, InfiniBand, optics, SW for DC fabrics (AI + non-AI)	**$39**	**$73**	gminsights.com
Back-end networking	AI cluster fabrics (Ethernet/IB) & optics tied to accelerator pods	**$20**	**$80**	650group.com, sdxcentral.com
Data movement / memory interconnect	CXL switches, controllers, memory expanders	**$1**	**$6**	gminsights.com
HPC (traditional market, AI chips)	On-prem HPC servers, storage, SW, services, plus cloud HPC	**$60**	**$87**	MarketsandMarkets
Hyperscalers (AI infrastructure CAPEX)	Big 5 cloud AI datacenter build (servers/GPUs, power, facilities)	**$300**	**$5200**	Morgan Stanley

Notes :

1. Most of the above segments may require production BI or specialized high-power HTOL
2. Most segments will have Silicon Photonics integrated by 2030
3. Scope overlap: AI training, inference as intersecting sets with some overlap with AI infrastructure

AI Infrastructure expenditures

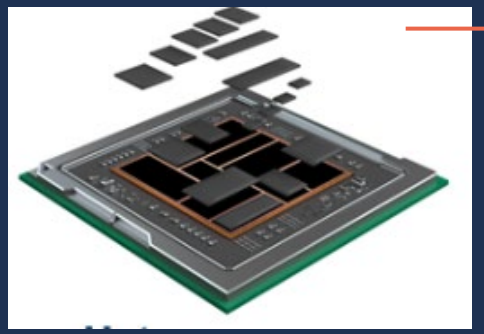


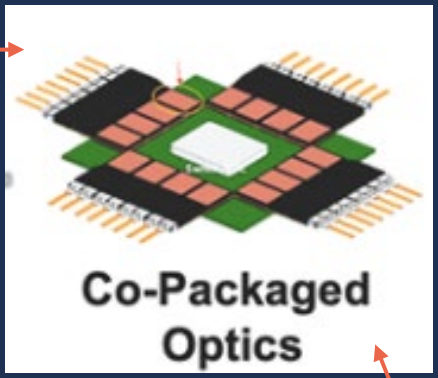
CAGR : 28.5%

Source : GrandView research

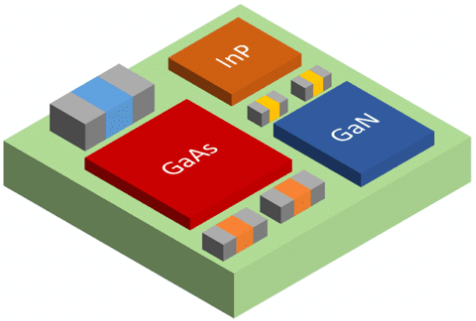
Year	Projected Global AI Infrastructure Capex	Key Trend / Driver
2023	**$35.4 – $36.4 Billion**	Build out of ChatGPT.
2024	**$45.5 – $46.2 Billion**	Scaling starts
2025	**$375 – $443 Billion**	Capacity Expansion
2026	**$660 – $700 Billion**	The SPRINT Year, be a part of it or loose it!
2027	**$629 – $800+ Billion**	Growth rates expected to moderate to ~17% as infrastructure matures.
2029	**$758 Billion**	Forecasted market size specifically for accelerated servers.
2030	**$1.7 Trillion**	Total data center infrastructure spending (AI-driven) projected by McKinsey.

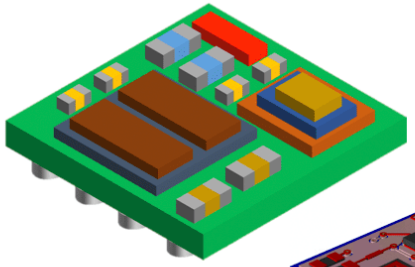
Evolution of HPC / HIR

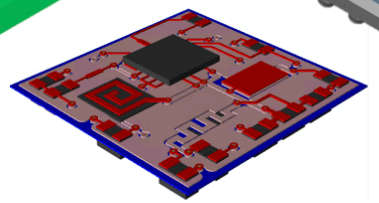


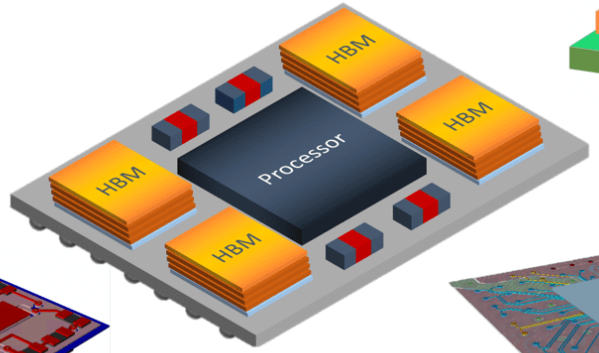


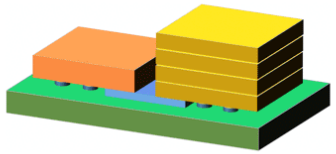
Heterogeneous Integration

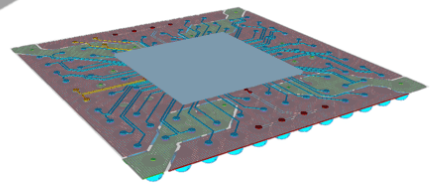


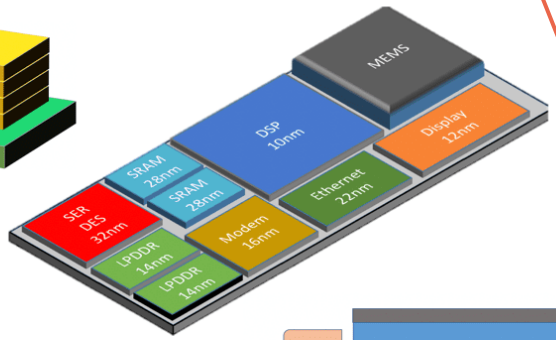


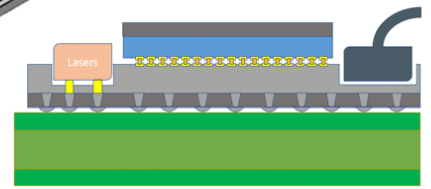










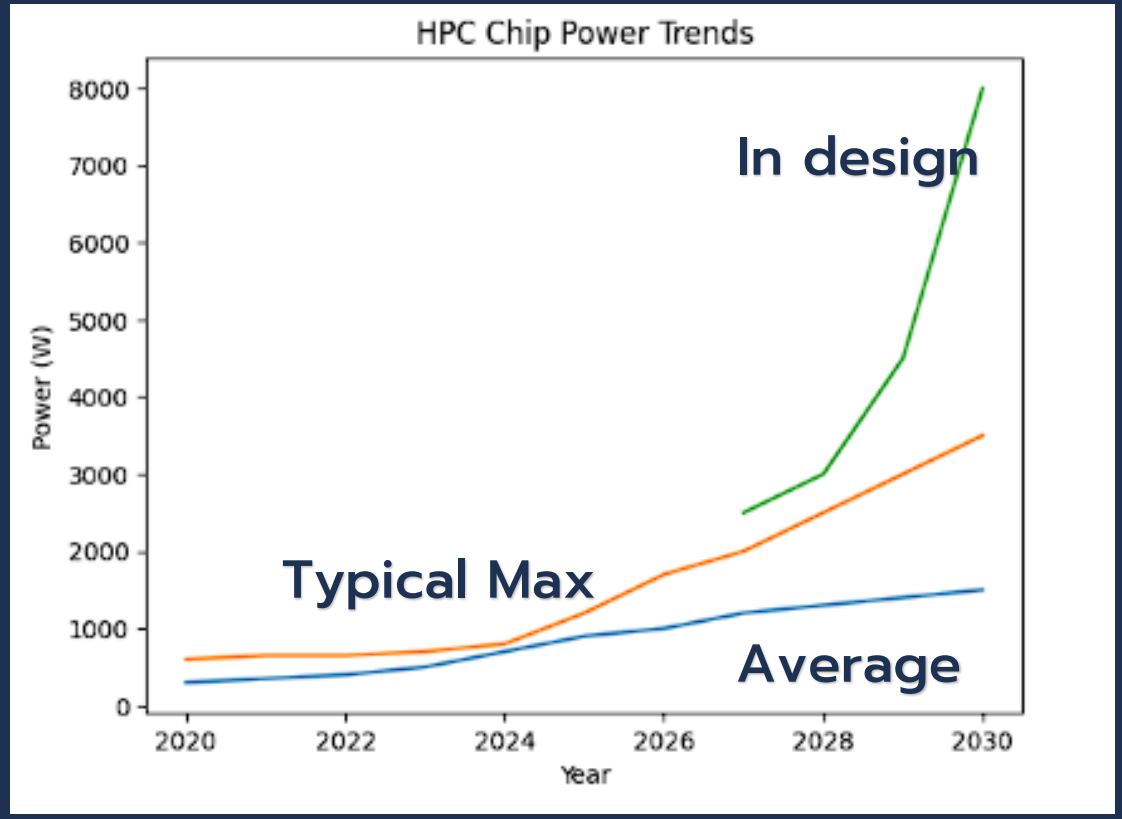


Multi-Chip Module (MCM)	System in Package (SiP)	RF Module	2.5D-IC (Silicon Interposer)	Embedded Bridges	High-Density RDL/FOWLP	3D-IC (Chip on Wafer)	Heterogeneous Integration (Disaggregated SoC) Photonics
1980	1990	1998	2005	2008	2014	2016	2018	Now

Why Burn-In is Suddenly Strategic?

Era/Year	Typical High-End CPU TDP	Typical High-End GPU TDP	Key Trend
Pre-2010	~119 W	~200 W	Serial processing; focus on clock speed.
2016-2018	~140–165 W	~250–300 W	Multi-core scaling becomes the standard.
2022-2024	~300–400 W	~450–700 W	Massively parallel architectures; 100+ CPU cores.
2025+ (est)	400 W+	700 W–1,000 W+	Liquid cooling transitions from niche to standard.

- **Data Center Socket Power Growth:** Mean power consumption per socket at full load increased by approximately **2.5x** from 2010 to 2022, rising from 119 W to 303 W with continued growth into 2030 (ie Intel Xeon / AMD EPYC)



- Power density doubling
- Thermal gradients intensifying
- Failure modes shifting
- Yield sensitivity increasing

Core idea: Burn-in is now part of product architecture, not just qualification.

Characteristics of HPC

➔ Power is now the dominant factor

	Typical HPC Chip Power Consumption (approx)			
Year	Typical	MAX power	in design	Notes / Basis
2020	300	600		Early HPC GPUs (e.g., A100 PCIe) and accelerators ≈ 200-500 W; average trending ~300 W.
2021	350	650		Successors to 2020 accelerators begin trending higher.
2022	400	650		Newer designs push up TDP even as efficiency per FLOP improves.
2023	500	700		Many deployed AI/HPC chips 500+ W.
2024	700	800		Leading accelerators like H100 rated up to ~700 W in SXM form factor.
2025	900	1200		Trends in upcoming chips (Gaudi 3 ~900 W; Blackwell ~1,200 W+ per GPU).
2026	1000	1700		Consensus industry direction shows power envelopes climbing as performance scales.
2027	1200	2000	2500	Continued push for performance in AI/HPC likely drives TDP up.
2028	1300	2500	3000	Efficiency gains might slow absolute growth but sustained upward trend.
2029	1400	3000	4500	Chip designs incorporate larger accelerators and larger memory.
2030	1500	3500	8000	Rough projection reflecting continued scaling of high-end chips.

- Between 2020–2025, Flagship AI accelerators have moved from ~400W to approaching 1000W
- From 2025–2030, if current architectural and performance drivers continue, projected average consumption could increase another ~60 %.
- This projection assumes *high-end designs* dominate HPC and AI workloads, as they historically have.

Sample Device Types Tested on the High-power BI System

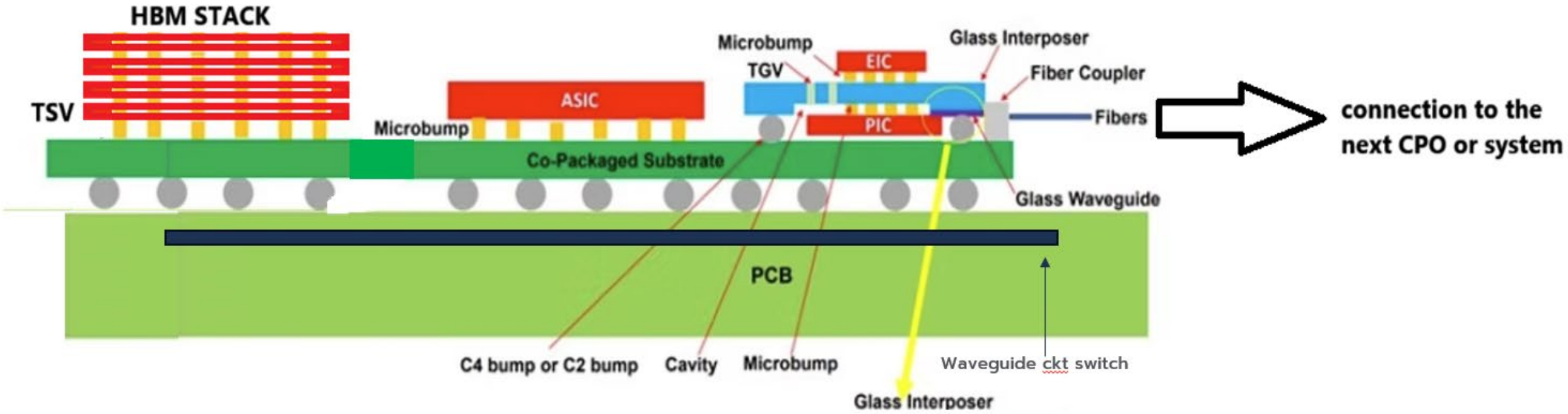
- Mission-mode dynamic stress profiles are difficult to replicate under traditional HTOL conditions.
- difficult to replicate true mode during HTOL/BI
- HTOL duration may need adjustment to compensate for lower stress acceleration.

Device List	Device Type / Application	package size (mm)	Power(W)
Device # 1	28nm Network Processor	55 x 55	130
Device # 2	28nm IP Test Chip	25x25	25
Device # 3	16nm GPU AI chip	25x35	35
Device # 4	28nm AI chip	19x19	12
Device # 5	16nm GPU AI chip	60x60	350
Device # 6	28nm Gigabit switch	57.5 x 57.5	200
Device # 7	16nm Network Router chip	62.5 x 62.5	550
Device # 8	16nm Router chip	66 x 56	330
Device # 9	7nm Processor	60 x 60	400
Device # 10	7nm Processor	60 x 60	700
Device # 11	7nm Processor	65 x 55	450
Device # 12	7nm AI chip	49 x 49	150
Device # 13	5nm Network processor	65 x 60	700
Device # 14	7nm AI chip	65 x 65	425
Device # 15	16nm GPU AI chip	50 x 55	300
Device # 16	Fiber Optic switch	49 x 49	125
Device # 17	7nm Network processor	60 x 60	450
Device # 18	7nm Network processor	60 x 60	450
Devcie # 19	3nm Network processor	90x90	1000
Device # 20	3nm Network processor	75 x 80	>800
In development			
Devcie # 21	3nm AI Chip	90x100	1200
Device # 22	3nm AI Chip	75 x 80	>900
Device # 23	3nm AI Chip	100x100	1500
Device # 24	2nm Network processor	110x120	1700
Many devices	2nm AI Chip	Bigger	More power

In queue

HPC's DNA Reimagined: Testing requirements

The end-goal is the full assembly including many elements – need test Production Burn-in!



1. Future PCB & interposers will have optical waveguides
2. Memory roadmaps are exploring optical IO to address bandwidth density constraints.

Co-Packaged Optics 1.0: Typical Integration Flow

Advanced Packaging with SiPh for CPO

Fiber Attach Assembly and Test

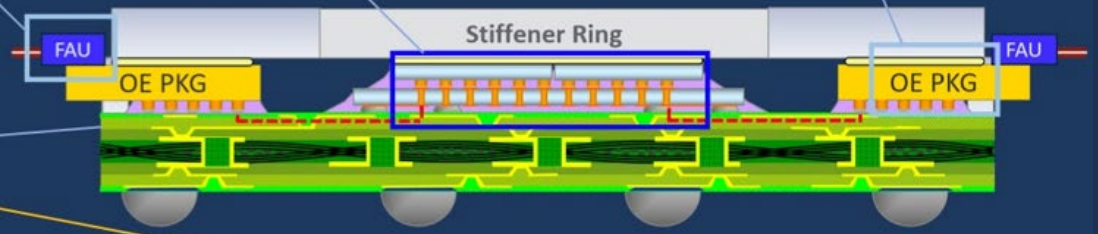
- Coupling approach (AA & PA)
- FAU detachability and reliability

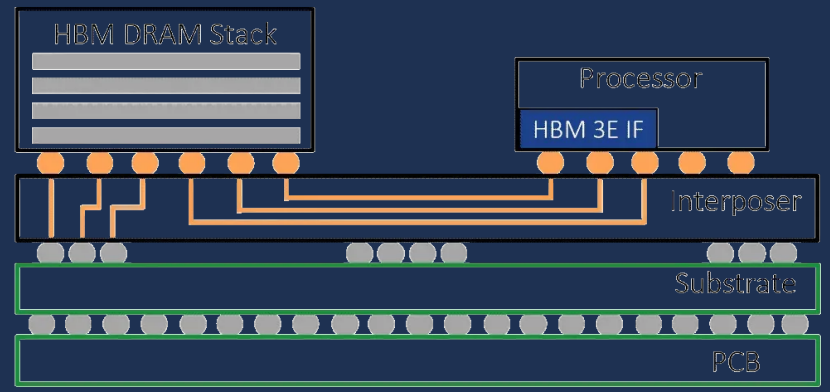
Chiplets Options

- Advanced HI packaging

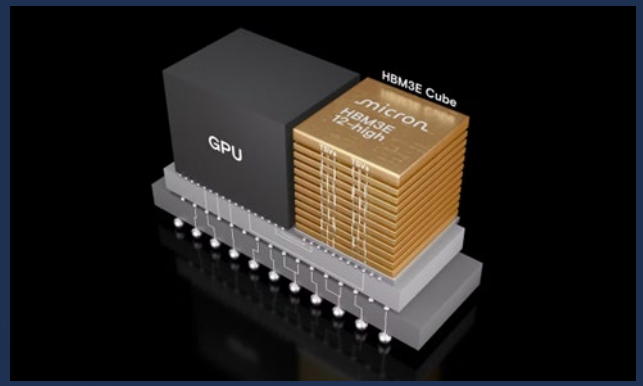
Optical Engine Solution

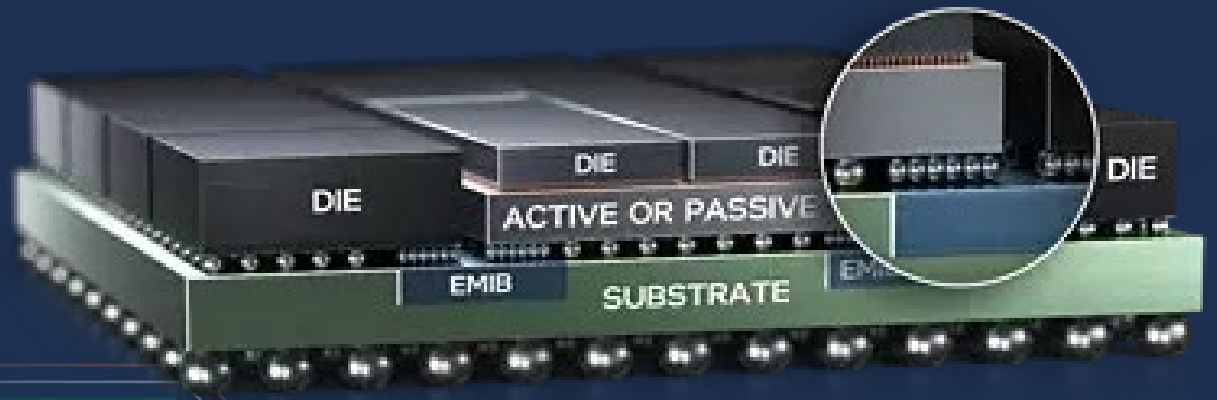
- EIC/PIC warpage optimized
- Test for KGOE & Optical facet quality





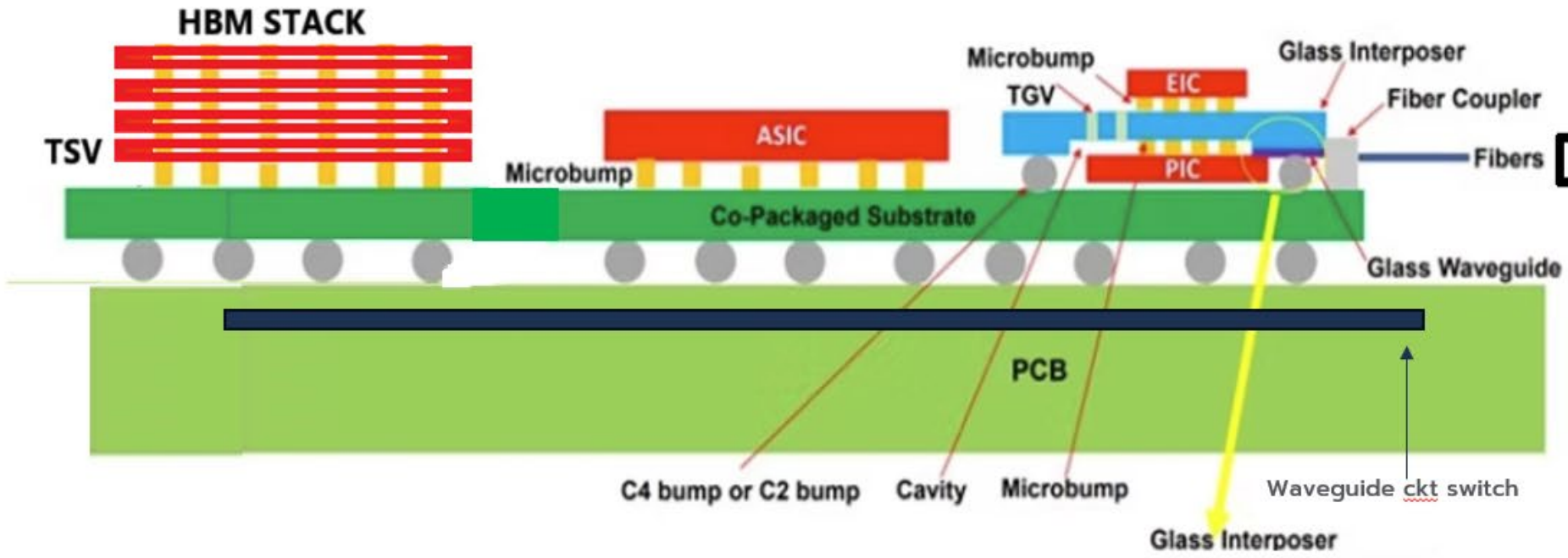
IF = HBM3E Interface





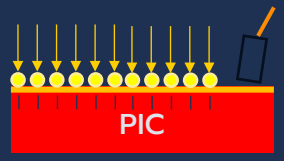
HPC's DNA Reimagined: Testing requirements

The end-goal is the full assembly including many elements – need test Production Burn-in!

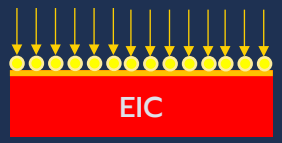


connection to the next CPO or system

Test 8, 9, 10
Final test, system level test, Burn-in

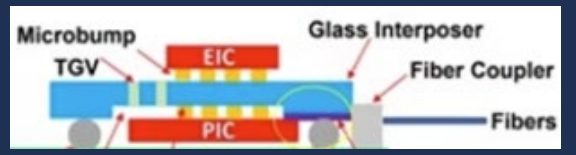


Test 1 & 2
PIC Wafer Testing
PIC Wafer Level Burn-in

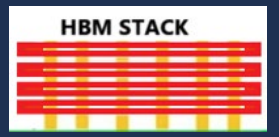


Test 3
EIC Wafer Testing

Test 4, 5
High-Power test/production Brun-In



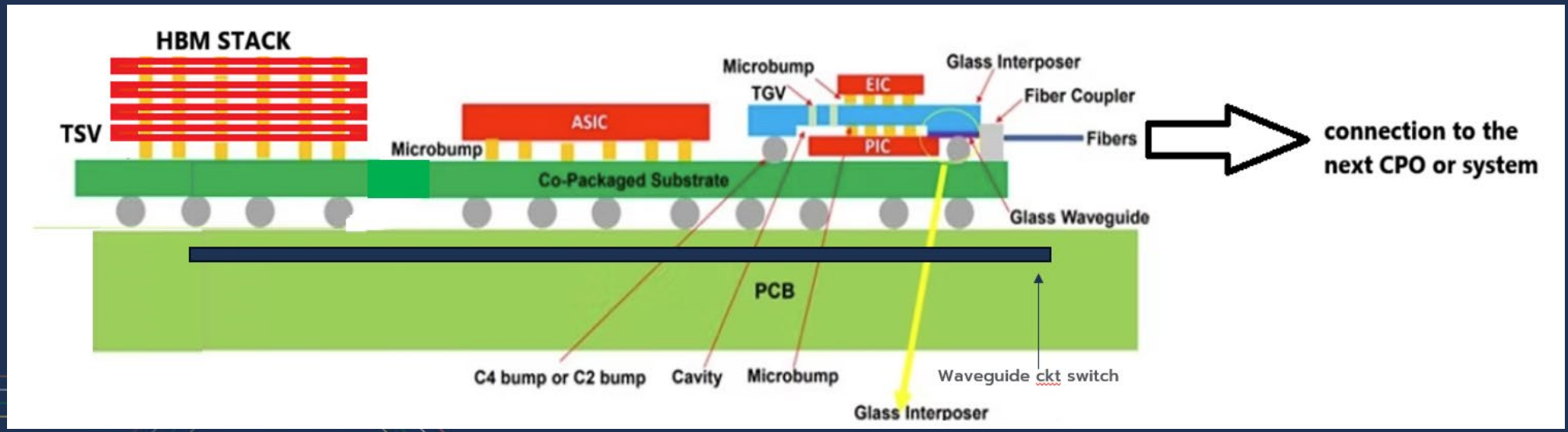
Test 5 & 6
Optical Engine Testing
OE Wafer Level Burn-in



Test 7
HBM3/4 test, Burn-in

The New Reliability Reality for 1 kW+ AI/HPC Packages

- HPC & AI Packages Have Entered the Kilowatt Class
  - Reliability is no longer a die-level issue — it's a thermal challenge across multiple die within the package; requirement for multizone TC
- Co-packaged (CP) and chiplet architectures (UCIe) introduce new failure modes:
  - Die-to-die interconnect degradation (UCIe lanes)
  - PDN and thermal–mechanical fatigue in multi-die stacks
  - Co-packaged optics/photonics sensitivity to thermal stress
  - Legacy JEDEC HTOL (die-only) testing misses package–system interactions
- Early screening is critical, as marginal die escapes can fail after expensive packaging.



Shift-Left Qualification Approach

1. Wafer-Level Burn-In (WLBI) → Known-Good-Die (KGD)
    - Stress die at wafer stage (T/Bias/AI-pattern vectors).
    - Catch latent defects early → feed KGD into assembly.
    - Reduces package scrap / rework, accelerates yield ramp.

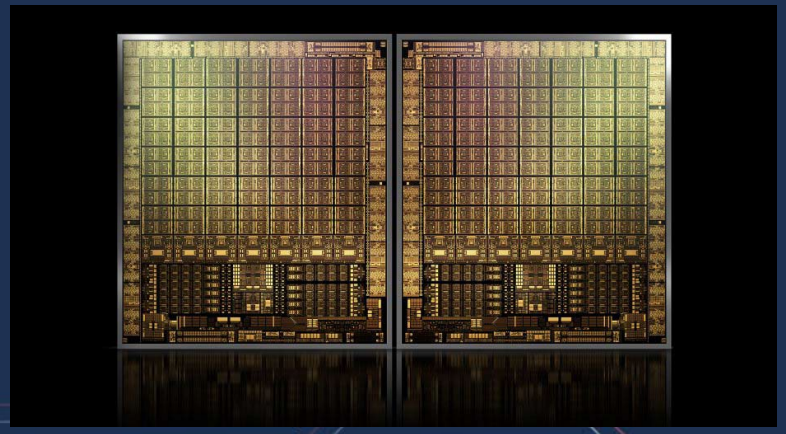
2. Package Level HTOL (System Co-Stress)
    - Include UCIe lane SI, PDN, and hotspot cycling
    - Run distributed functional workload patterns throughout the chip
    - Combine with mechanical stress?

3. Data Integration & HIR Alignment
    - Align test plans to Heterogeneous Integration Roadmap (HIR)
    - evaluate WLBI for HTOL and production Burn-in leading to SLBI and SLT

➔ Start planning early with DFT

Action Plan: Industrialize the Shift-Left Qualification Flow

- Implement WLBI infrastructure
  - For high power devices, implement distributed stress across high power die for HTOL and production BI
- Define unified HTOL test matrix for chiplet / CPO packages.
- Embed hooks for PDN / thermal monitoring.
- Establish an ecosystem for TV-based learning pre-silicon to improve design robustness and predict reliability for thermals



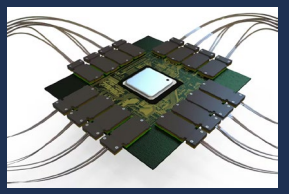
Fermi Streaming Multiprocessor (SM)

Test Components of “Shift Left” for CPO

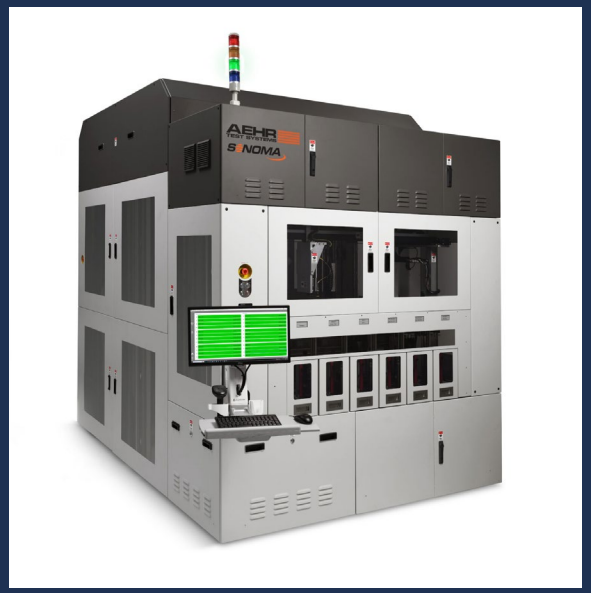
1. ATE wafer test
    1. All components of CPO
2. WLBI (Wafer Level Burn-In)
    1. HTOL
    2. Production Burn-In
3. Optical testing
    1. PIC
    2. OE (Optical Engine)
    3. Sub assembly (O/E test)
    4. Production Burn-In of OE
4. Stacked die testing
    1. HBM / OE / CoW or CoWoS
    2. Production BI
5. Assembled module test
    1. ATE
    2. SLT
    3. Production Burn In











Burn-In Across the Manufacturing Flow

From Die to Package to System: Burn-In is Shifting Left and Scaling Up

1 .Wafer-Level Burn-In (WLBI)

- Stress before packaging → removes weak die early; enables KGD
- Thermal density manageable (<20–50W/die); ideal for chiplet ecosystems (UCIe, HBM, PIC)

Package-Level Burn-In (PLBI) for high power

- Captures package-interaction failures; hotspot formation at compute die, HBM, OE
- Requires multi-zone thermal control; chamber BI insufficient

System-Level Burn-In (SLBI/SLT)

- Required for 600W–2000W devices & liquid-cooled modules; board & VRM coverage

How Burn-In is changing: Air → Water → Multi-Zone

➔ Air-Cooled Era (≤200W GPU/CPU)

- Uniform temperature assumptions; chamber BI sufficient with individual cooling sufficient
- single-zone thermal profile Failures dominated by die-level defects; minimal hotspot interaction

➔ Water-Cooled Era (200W–1kW accelerators)

- Direct liquid cooling to maintain Tj; steep gradients between compute and HBM
- Chamber BI no longer maps to real system thermals; package interactions start to dominate; formation of hot spots due to design, socket, etc

➔ Multi-Zone Thermal Era (1–2kW+ AI/HPC)

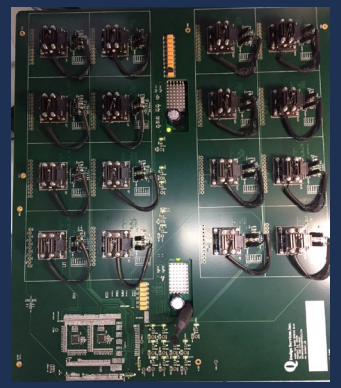
- Chiplets, HBM, and optical engines each create unique thermal zones; ΔT >50°C
- Needs independent thermal actuation, load shaping, and real-time thermal feedback
- Single-zone BI over/under-stresses → yield loss and escapes
- Cooling from the bottom of the device to provide thermal control of solder ball to silicon

Wafer Level Burn-in (WLBI - HTOL) “the transition”

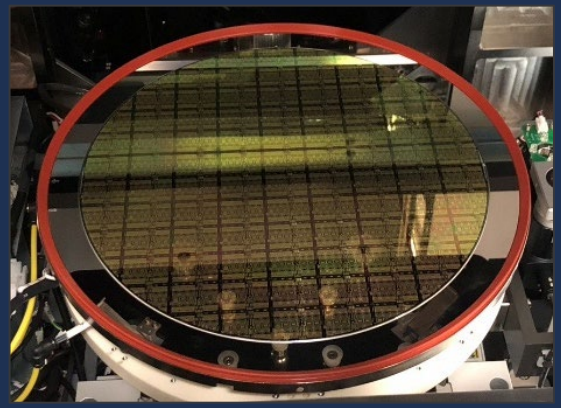
1. Important for chiplets that don’t go into packages anyway
2. High cost of package qual: Building boards, buying sockets, ATE read point testing
3. Ability to run three lots one on wafer, ATE read points a matter of running one wafer. (assuming <15W / die)

Current solutions provide ~3500 Watts per wafer

➔ For 240 units ➔ 15watts / die
➔ For 160 units ➔ 22watts / die
➔ For 80 units ➔ 45watts / die



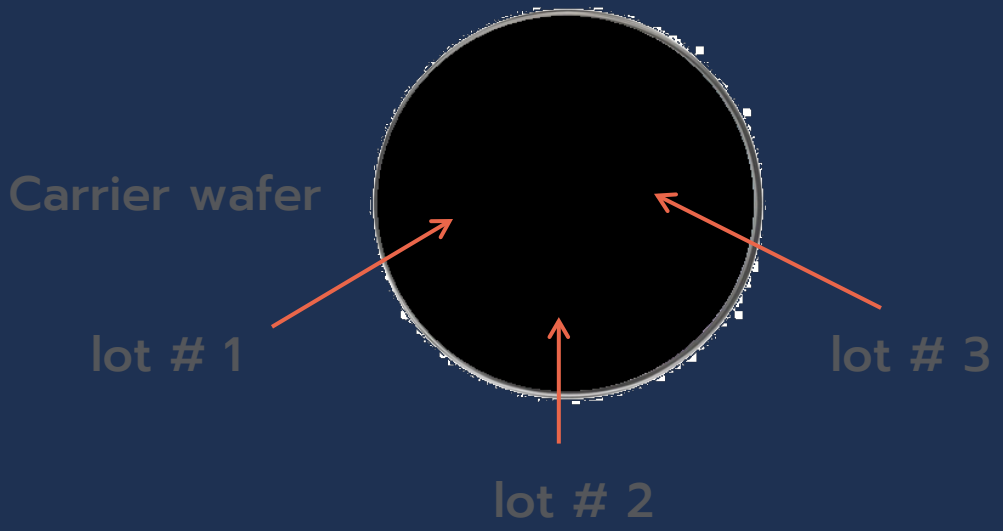
Traditional BIB

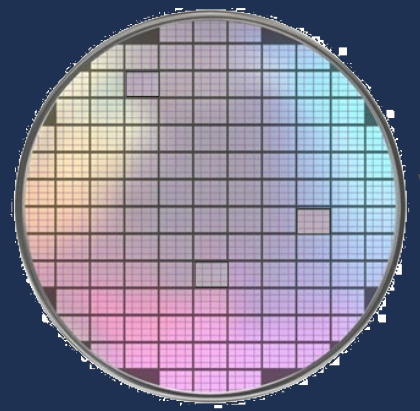


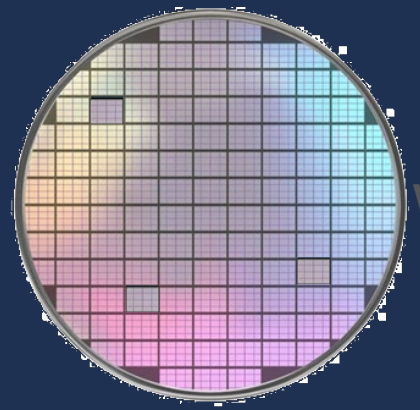
300mm Wafer on
Aehr FOX-XP WaferPak “Chuck”

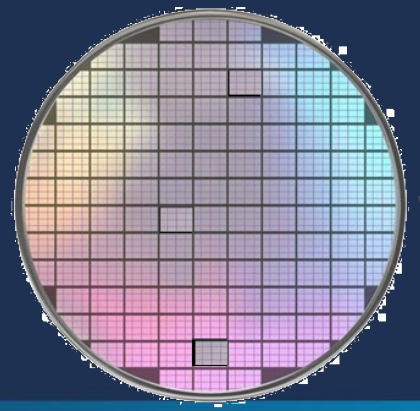
Transfer of die

- Running three lots on one carrier wafer




Wafer lot # 1


Wafer lot # 2

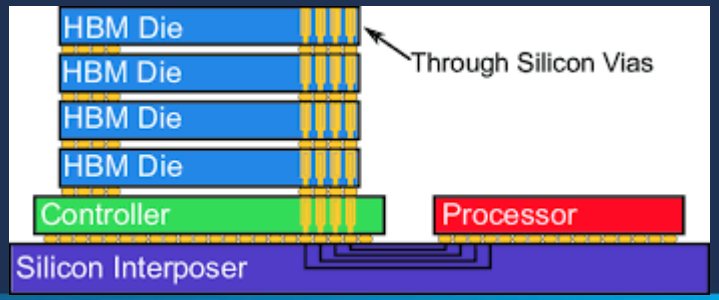

Wafer lot # 3

Wafer Level Production (WLBI-PBI) “the other transition”

Strategic Shift : Burn-in is moving upstream — from system-level screening to die- and component-level qualification before final packaging.

Why It’s Now Mandatory:
- Package value: $20K–$40K+ per module
- Chiplet + 8 HBM stack + optics = yield multiplication risk
  - Example : 80% chiplet, 95% HBM, and 60% optics yield,
  - (.8 x 0.95^8 x .6) = 31.84% yield (illustrative only)
- Field failure = full module scrap
- Hyperscalers demanding ultra-low FIT

1 FIT → 1 Failure In Time in 10^9 hours

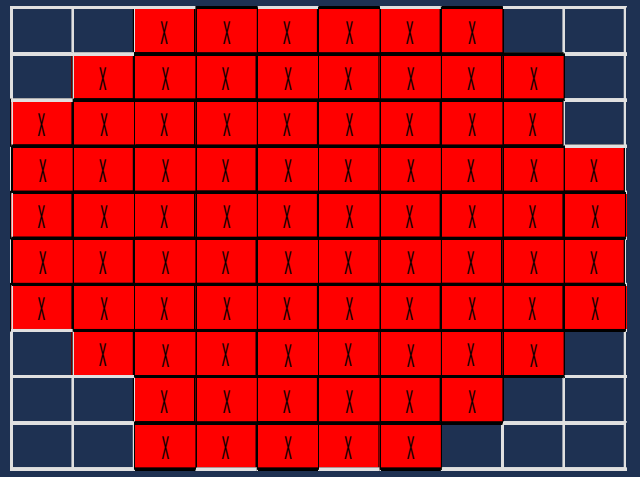


Wafer Level Production (WLBI-PBI) “the other transition”

1. KGD is -critical for chiplet-based compute dies in HPC and CPO packages
2. Multiple components require production-grade burn-in (BI)
   - ➔ Compute DIE (GPU/CPU/cache)
   - ➔ Chiplets (compute I/O / OE - PICs)
   - ➔ HBM (before stacking and placing into package)
   - ➔ Integrated lasers
   - ➔ High speed SerDes chiplet

➔ What if the total wafer thermal power exceeds 3500Watts?

Production BI : high power devices



Example : compute die

a. 350W production BI max power
b. 80 die per wafer
c. 2-hour Production BI time
d. Test made up of MBIST, LBIST, Serdes loop back, etc

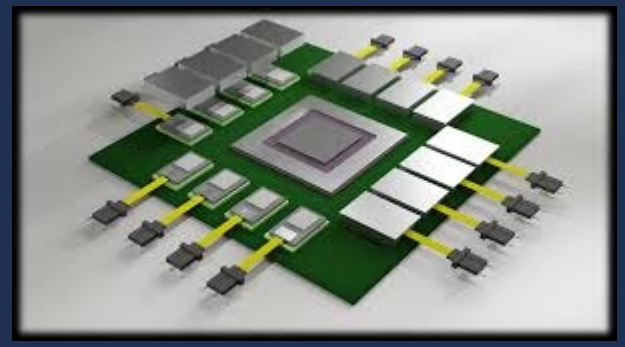
In this case, 350W x 80 devices = 28KW (3500W limit!)

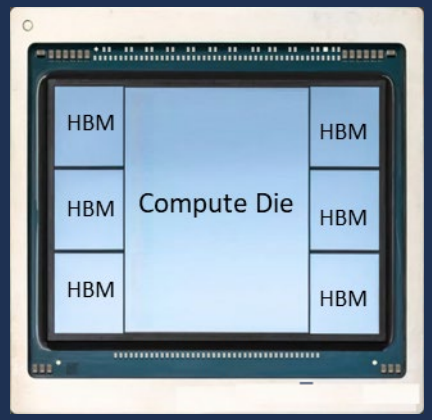
Options

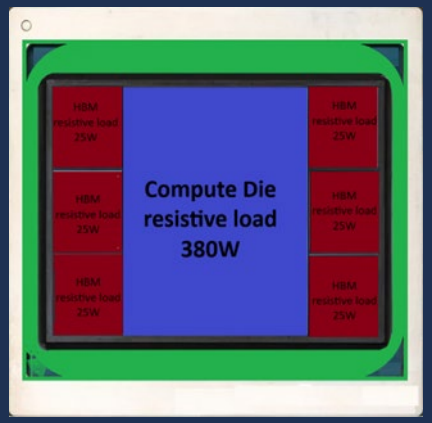
1. Test 10 die at a time, requiring 16 hours total production BI time vs 2hrs
2. Test 40 die at a time, work with DFT to test to reduce vectors to exercise only 25% of the cores/memory at a time, requiring a two pass test but 25% of the power per die

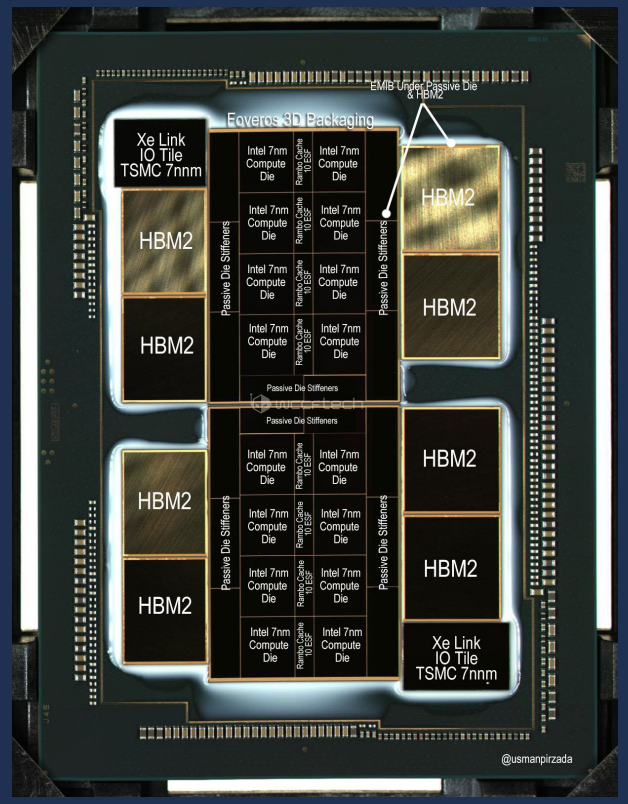
PLBI for HTOL and production BI for HPC

DNA consists of:
1. Compute die(s)
2. HBM (stacks)
3. Silicon interposer
4. Heat spreader
5. OE
6. Laser
7. FAU
8. waveguides









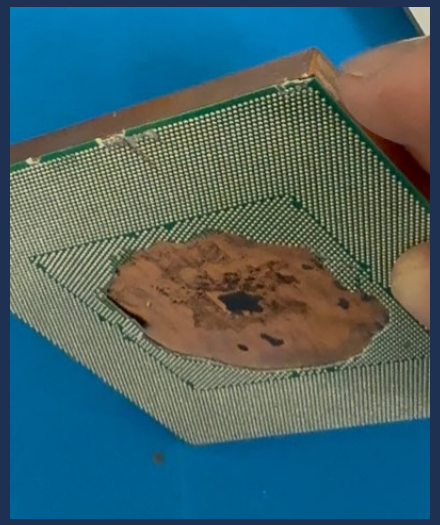
Example ➔
530W device

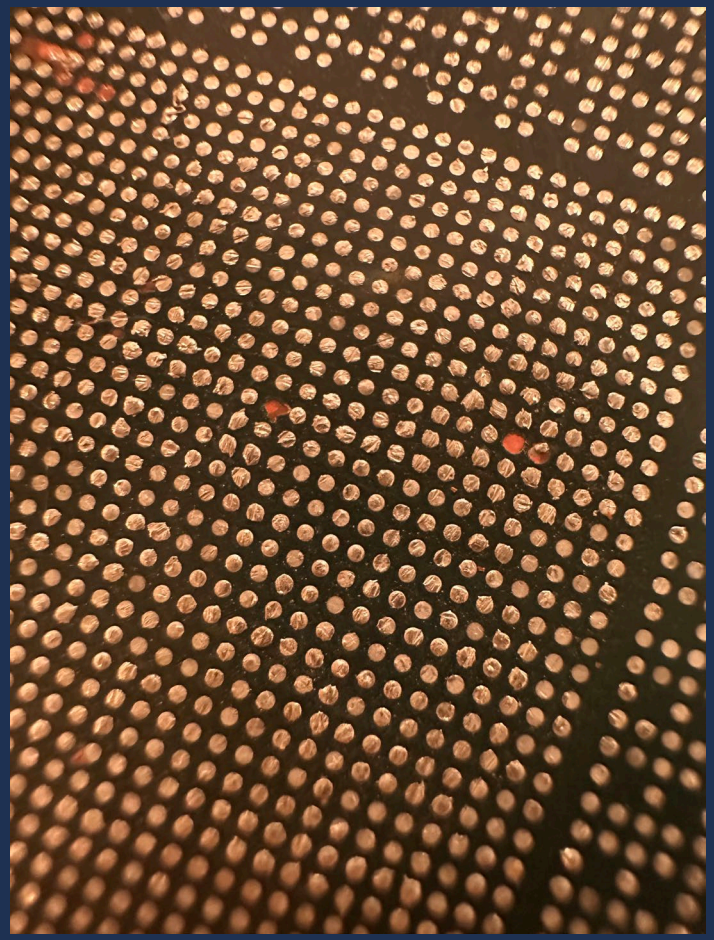
New requirements for HPC HTOL / PBI

1. Multi-zone Thermal control (individual thermal regulation)
2. Hot spot management (control di/dt)
3. Bottom side cooling
4. Mechanical loading on part (pressure, package variations)
5. Interface material (high-K TIM)

Thermal runaway causing:

- Solder ball fatigue
- Interconnect fatigue/discontinuities





Disciplined execution across many fronts.

(3 examples)

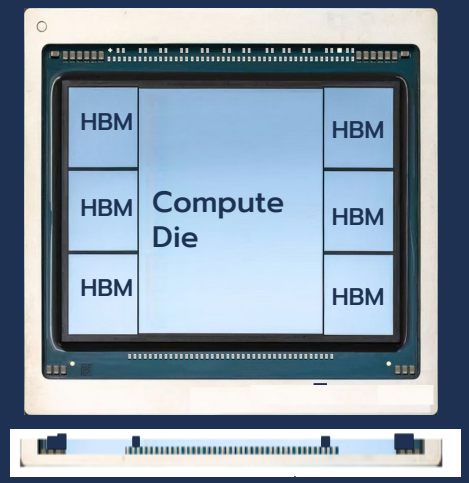
1. Build a representative test vehicle (TV) before first silicon
   - Not a simulation shortcut, but real thermals in a realistic package, operating under true system conditions.
   - If you want credible data, measure it in the form factor that actually ships in first article HW (cooling infrastructure, socket, etc)

2. Map the full thermal profile of the chip.
   - Not just peak temperature — but gradients, hotspots, transient behavior, and workload-dependent variation.
   - Understand how heat moves across the die, through the package, and into the system. Thermal truth lives in the details
   - DFT simulation required

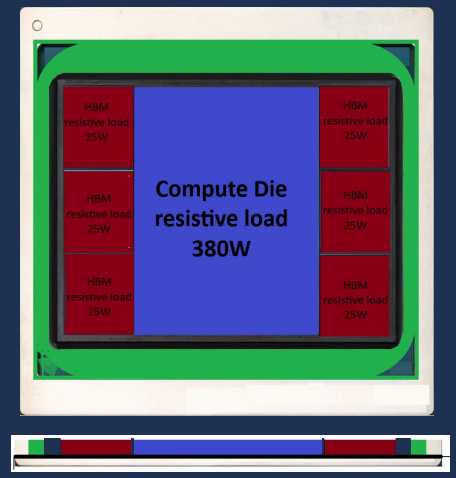
3. Implement preemptive power management.
   - Don’t wait for thermal runaway to react.
   - Design the system to anticipate load spikes, dynamically balance performance and power, and manage heat before it becomes a limiter.

Example 1 : 530 Watt HPC AI die

Actual device

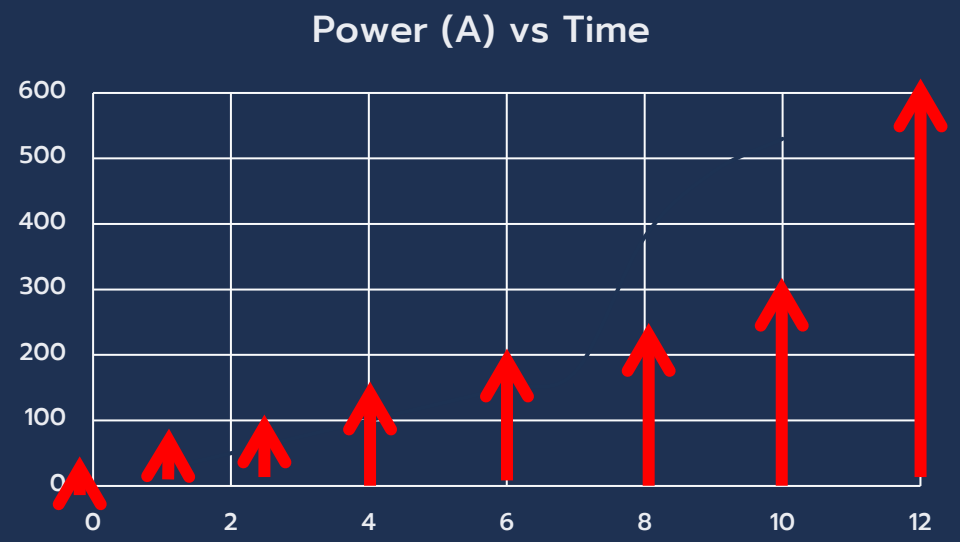


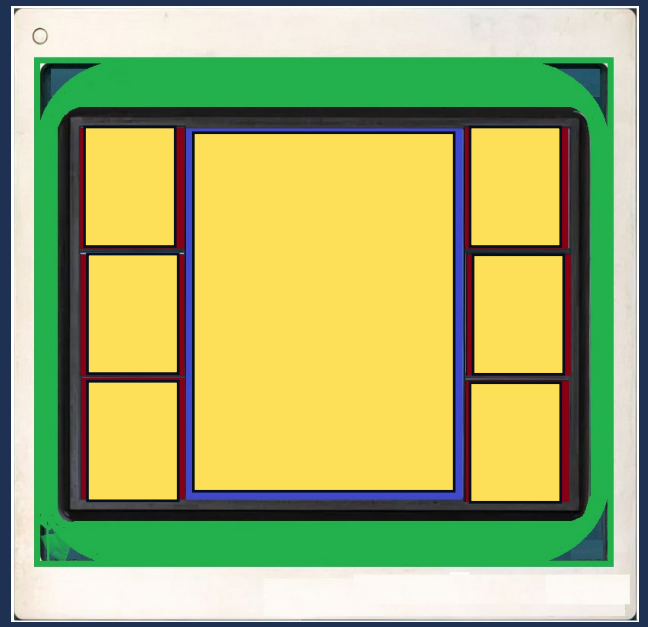
TV (test Vehicle)

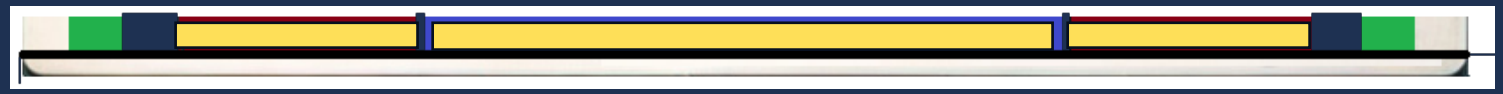


PLBI for HTOL and production for HPC

Can cycle power like the actual device

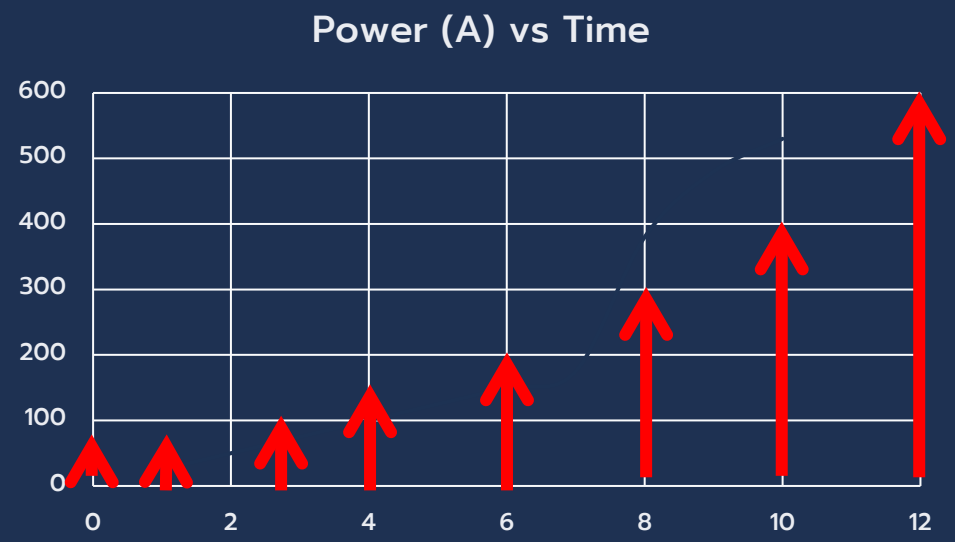


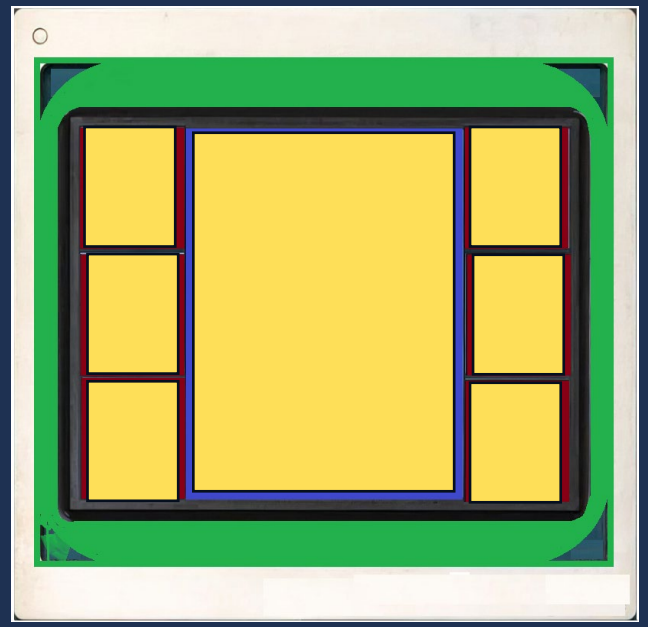


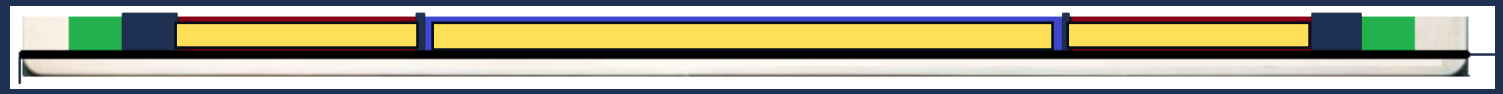


PLBI for HTOL and production for HPC

Can cycle power like the actual device

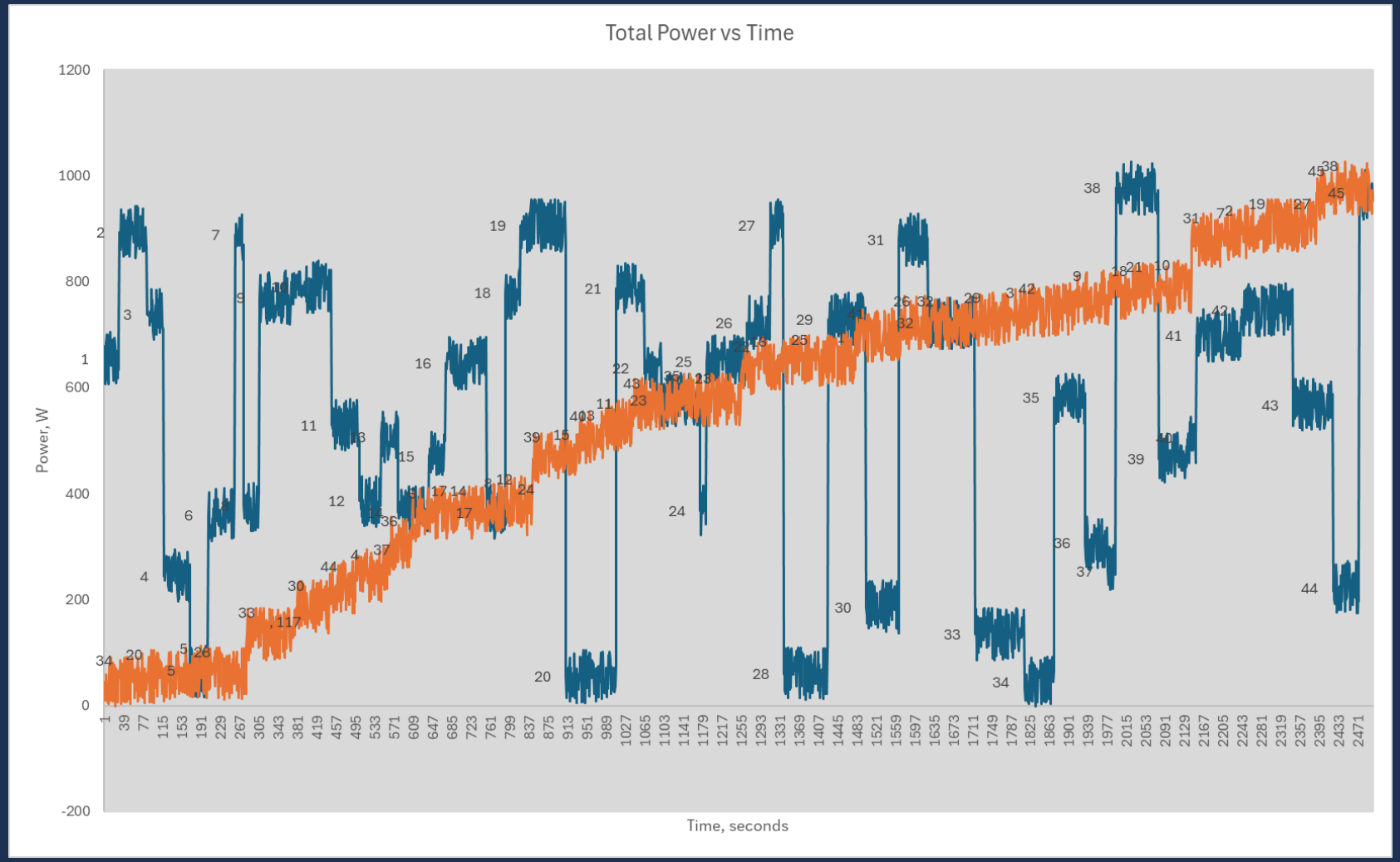






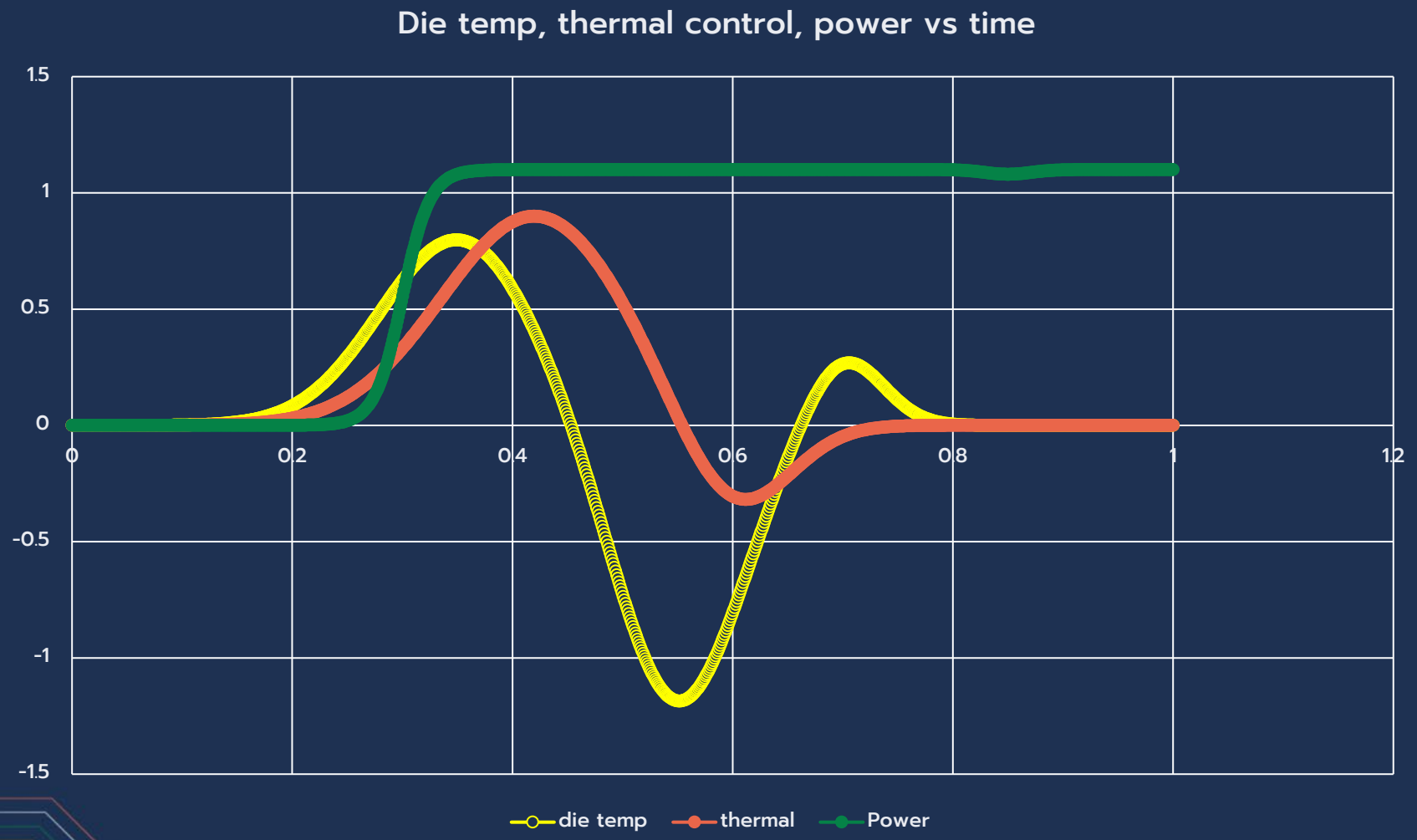
Example # 2 use AI to (reduce di/dt)

1. Dual-Core V2 Evolution: Enhanced architecture with expanded HBM integration.
2. Rapid Validation Cycle: 45-test comprehensive program in <40 minutes.
3. AI-Optimized Power Stability: Intelligent vector reordering for uniform thermal loads.



Example # 3: Use AI to control thermal preemptively

1. Thermal signature is already known in a test program that repeats
2. Use AI to figure out when to start cooling to avoid overshoot



The Kilowatt Era Has Arrived!

AI didn't just increase performance. It changed the physics of reliability.

- 300W → 900W → 1kW+ packages
- Chiplets + HBM + CPO multiplying yield risk
- Thermal gradients now define failure mechanisms
- A single escape = $20K-40K module loss + AI module at risk

Legacy qualification models were built for monolithic dies.

The future is heterogeneous, optically interconnected, and thermally constrained.

Traditional die-level HTOL alone does not capture heterogeneous package interactions.

The Shift Is Strategic

Burn-In is no longer:

- A screen
- A filter
- A qualification checkbox

Burn-In is now:

-An architectural requirement
-A yield protection strategy (multiplier)
-A competitive advantage

Reliability data must close the loop into design

The Companies That Win the AI Decade Will:

- Design for stress.
- Industrialize wafer-level and package-level burn-in.
- Control multi-zone thermals at the kilowatt scale.
- Integrate optics, compute, and 3D memory with a zero-escape mindset.

HPC → AI was the inflection point.

Kilowatt reliability is the new battlefield

Thank you

Back up slides

HPC – The Inflection Point broken out by Segment (in Billions)

Segment	Incl	Today (2024/2025)	Outlook (2030)	Source
AI Accelerators	GPUs + AI ASICs/ASSPs for cloud/DC (training & inference)	**$207**	**$286**	Omdia's latest forecast for AI processors in cloud & DC; $123B in 2024 → $207B 2025 → $286B 2030. [omdia.tech...nforma.com]
AI Inference	HW & platforms serving inference workloads DC/cloud/edge	**$160**	**$255**	MarketsandMarkets base case. Grand View puts 2024 at $97B and 2030 at $254B—essentially the same magnitude.
AI training	Training portion of DC GPU/accelerator spend	**$74**	**$140**	Derived by applying **~62% training share** to the DC-GPU market ($120B in 2025; $228B in 2030). Share from FMI (training 61.7%); DC-GPU totals from MarketsandMarkets.
Data Center networking	Ethernet, InfiniBand, optics, SW for DC fabrics (AI + non-AI)	**$39**	**$73**	Nextmsc global DC networking. GMI shows similar trajectory (mid-$20Bs in 2025 rising strongly this decade). [nextmsc.com], [gminsights.com]
Back-end networking	AI cluster fabrics (Ethernet/IB) & optics tied to accelerator pods	**$20**	**$80**	650 Group pegs **~$20B in 2025**; Dell'Oro expects **~$80B** DC switch sales for AI back-ends over the "next five years" and Ethernet overtaking IB. [650group.com], [sdxcentral.com]
Data movement / memory interconnect	CXL switches, controllers, memory expanders	**$1**	**$6**	Forecasts vary by scope: GMI (components) vs. Strategic Market Research (broader CXL components). Range shown. [gminsights.com], [strategicm...search.com]
HPC (traditional market, AI chips)	On-prem HPC servers, storage, SW, services, plus cloud HPC	**$60**	**$87**	Hyperion (HPC/AI/quantum/cloud combined) shows **$60B in 2024** and **>$100B by 2028**
Hyperscalors (AI infrastructure CAPEX)	Big 5 cloud AI datacenter build (servers/GPUs, power, facilities)	**$300**	**$5200**	Morgan Stanley (via Yahoo) >$300B 2025; MUFG/IEEE ComSoc puts **2026 >$600B**; McKinsey baseline shows **$5.2T** *AI* DC capex through 2030 (part of **$6.7T** total DC capex). **CapEx, not revenue.**

Notes :

1. Most of the above segments may require production BI or specialized high-power HTOL
2. Most segments will have Silicon Photonics integrated by 2030
3. Scope overlap: AI training, inference as intersecting sets with some overlap with AI infrastructure

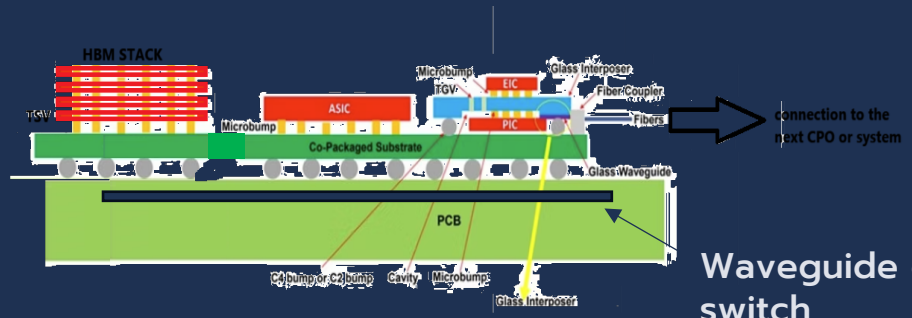
Industry Changes Needed to Test These Parts?

Optical Engine

- ✓ Precise positioning of fiber coupler.
- ✓ Typically, optical connector on opposite side as electrical connections (probes)
- ✓ Fuse blowing
- ✓ Temperature Control.
- ✓ Vibration control (ground-floor location)
- ✓ Clean-Room environment

➔ Critical step for Known-Good-Chiplet

Optically Coupled Hyper-Scalar



Waveguide ckt switch

- ✓ Precise positioning of multiple fiber couplers.
- ✓ Optical connections on four sides, electrical connection from bottom, and cooling from top.
- ✓ Vibration control (ground-floor location)
- ✓ Clean-Room environment

Wafer Level production Burn-In not possible?

Modern AI accelerators are starting to exceed 800–1200W package power hard to realistically stress at wafer level.

Consolidated Decision Matrix

Device Characteristics	WLBI	PLBI	SLBI/SLT
<450W monolithic die chiplet, OE or die module, HBM	✅	possible	❌
450–600W monolithic	⚠	✅	Optional
2.5D + HBM	❌	✅	⚠
600-2000W package	❌	possible	✅
Liquid cooled in field	❌	possible	✅
Chiplet mesh fabric	❌	✅	⚠
Hyperscaler critical	❌	⚠	✅

❌ Not viable ⚠ Limited realism ✅ Best option

COPYRIGHT NOTICE

www.testconx.org