WHITEPAPER · SEMICONDUCTORS & MATERIALS

Advanced Packaging & Materials for AI Accelerators

A Thermal & Yield Study

An engineering examination of 2.5D and 3D integration, chiplet disaggregation, and the thermal and materials decisions that now set the performance ceiling for AI accelerators — with a transparent, illustrative thermal-and-yield model.

AUTHORS
Root Digit Advanced Hardware Lab

PUBLISHED
2026

VERSION
1.0

CLASSIFICATION
Public

rootdigit.com

WHITEPAPER

Abstract & Contents

ABSTRACT

For five decades, performance gains in computing were delivered chiefly by shrinking the transistor. That engine is slowing: lithographic, power, and economic limits now constrain how much a single monolithic die can grow or improve. For AI accelerators — which demand enormous compute, vast memory bandwidth, and ever-higher power in a fixed footprint — the locus of innovation has shifted from the transistor to the *package*. Advanced packaging, 2.5D silicon interposers, 3D die stacking, chiplet disaggregation, and high-bandwidth memory integration, is how the industry continues to scale system performance. This whitepaper examines the engineering of that transition through two lenses that determine whether an accelerator can be built and operated profitably: *thermal extraction* and *manufacturing yield*. We describe the packaging and materials landscape, define the relevant metrics, present a transparent illustrative model of junction temperature and die-area yield, and report modeled results. The findings indicate that thermal resistance, not transistor count, increasingly limits sustained accelerator performance, and that chiplet disaggregation can recover yield that a large monolithic die would surrender to defect statistics — at the cost of new interconnect, thermal, and warpage challenges.

KEY FINDINGS AT A GLANCE

- Package-level thermal resistance, not raw transistor count, increasingly sets the ceiling on sustained accelerator throughput at high power.
- Yield falls steeply with die area under a Murphy defect model; splitting a large die into chiplets recovers a substantial fraction of that lost yield.
- Direct-to-die and near-junction cooling, paired with low-resistance thermal interface materials, can cut junction temperature by tens of degrees versus a conventional lidded stack.
- 2.5D and 3D integration shorten die-to-die links, raising effective bandwidth per watt — but concentrate heat and introduce thermomechanical warpage that the package must absorb.

Contents

SECTION ONE

1. Introduction

For most of the history of the integrated circuit, the path to a faster chip ran through a smaller transistor. That path is narrowing — and for the accelerators that train and serve modern AI, the package has quietly become the new performance frontier.

The observation popularly known as Moore's Law described a roughly biennial doubling of transistor density, and for decades it was accompanied by Dennard scaling, which let each new, smaller transistor switch faster while consuming less power. Together they delivered compounding, almost effortless performance growth: designers could rely on the next process node to make the same architecture faster and cheaper. Both of those tailwinds have weakened. Dennard scaling broke down in the mid-2000s as leakage current stopped falling with dimension, ending the era of free frequency gains and ushering in the multicore turn. Density scaling continues, but each node is now slower to arrive, far more expensive to develop, and yields ever-smaller per-transistor improvements in speed and energy.

This matters acutely for AI accelerators. Training and inference workloads are bound by two resources above all: raw arithmetic throughput and memory bandwidth. Both have grown faster than monolithic silicon can comfortably supply. A modern accelerator may dissipate several hundred to over a kilowatt of power, must feed its compute fabric from terabytes per second of memory, and is increasingly constrained not by how many transistors it can hold but by how large a single defect-free die can economically be manufactured — and how much heat can be pulled out of it.

Three pressures therefore converge on the package. First, the **reticle limit**: lithography can pattern only a finite area in a single exposure, capping the size of a monolithic die at roughly 800 mm^2. Second, **yield economics**: defect density makes very large dies disproportionately expensive, because the probability that a die contains a killer defect rises with its area. Third, **memory bandwidth**: getting enough data into the compute fabric requires placing high-bandwidth memory physically adjacent to the logic, across thousands of short, fast wires that a conventional printed circuit board cannot route.

Advanced packaging answers all three. By integrating multiple smaller dies, called chiplets, on a shared high-density substrate, designers escape the reticle limit, improve yield by building from smaller and more economically manufacturable pieces, and place memory stacks within millimetres of the logic for enormous bandwidth at modest energy per bit. The result is that an accelerator's ceiling is now set as much by packaging and thermal engineering as by the transistor itself.

This whitepaper examines that transition through the two constraints that most determine buildability and operating economics: **thermal extraction** — how much heat can be removed before the silicon throttles — and **yield** — how many good devices emerge from a wafer. We survey the packaging and materials landscape (Sections 2 to 4), examine reliability (Section 5), present a transparent and explicitly illustrative model (Section 6), report modeled findings (Section 7), and close with discussion, limitations, recommendations, and conclusions (Section 8). The figures throughout are modeled from representative physical pa-

rameters rather than measured from any single product; Section 6 states every assumption so the reader can substitute their own.

SECTION TWO

2. Advanced Packaging: 2.5D, 3D & Chiplets

Advanced packaging is the discipline of building a system out of several dies that behave, electrically and thermally, almost as one. The choices made here — how dies are arranged, on what substrate, and over which interconnect — fix the bandwidth, the power, and much of the heat the package must later shed.

The starting point is **die disaggregation**: rather than fabricating one enormous monolithic die, the design is partitioned into multiple smaller dies, or **chiplets**, each potentially built on the process node best suited to its function. Compute logic can use the most advanced and expensive node; analogue, I/O, and SRAM blocks, which scale poorly and cheaply, can use a mature node. The chiplets are then reassembled in the package. This decoupling of design from a single monolithic die is the organising idea behind most modern high-performance accelerators.

Two integration geometries dominate. In **2.5D integration**, multiple dies sit side by side on a **silicon interposer** — a passive slab of silicon carrying tens of thousands of fine wires and through-silicon vias (TSVs) that route signals between dies far more densely than any organic substrate could. Foundry platforms of this style (for example, chip-on-wafer-on-substrate, or CoWoS-class processes) are the workhorse of high-bandwidth-memory accelerators. In **3D integration**, dies are stacked vertically and bonded face-to-face, with TSVs or direct copper-to-copper hybrid bonds carrying signals through the stack. 3D shortens links to almost nothing and maximises density, but it stacks heat sources on top of one another — the central thermal problem of Section 3.

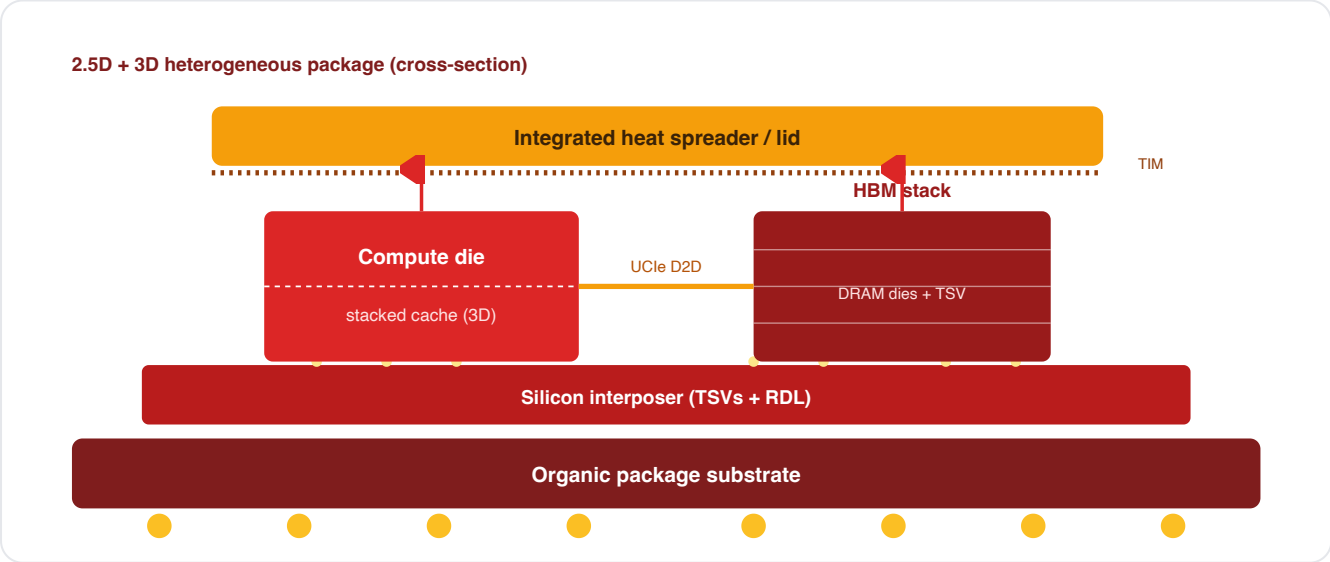

Figure 2.1 — A representative 2.5D + 3D accelerator package: compute die and HBM stack sit on a silicon interposer over an organic substrate, linked die-to-die and capped by a heat spreader through a thermal interface material.

Tying chiplets together demands a **die-to-die interconnect** that behaves like an on-die wire rather than a board trace. The industry has converged on open standards for this, most prominently **UCIe** (Universal Chiplet Interconnect Express), which defines the physical layer, protocol, and packaging assumptions so chiplets from different sources can interoperate. A good die-to-die link delivers terabytes per second across the package boundary at energy costs measured in sub-picojoule per bit — orders of magnitude better than crossing a board — which is precisely what makes disaggregation worthwhile.

High-bandwidth memory (HBM) is the other pillar. HBM is itself a 3D structure: a stack of DRAM dies bonded over TSVs and placed beside the compute die on the interposer. Its wide, short interface delivers the terabytes-per-second feed that AI workloads demand, at far better energy per bit than conventional graphics memory routed across a board. The consequence, though, is that memory and logic now share the same thermal envelope, and HBM is temperature-sensitive — a coupling that makes the thermal design inseparable from the packaging choice.

Packaging approaches compared

Approach	Interconnect density	Heat handling	Relative cost	Typical use
Monolithic die	On-die (highest)	Single plane, manageable	Low per part, poor at large area	Small / mid logic
Organic 2.xD / fan-out	Moderate	Good (planar)	Moderate	Cost-sensitive multi-die
2.5D silicon interposer	Very high	Good (planar, large)	High	HBM accelerators
3D stacked / hybrid bond	Highest (vertical)	Hardest (stacked heat)	Highest	Cache / memory-on-logic

The strategic takeaway is that packaging is now a first-order architectural decision. The geometry chosen — planar 2.5D, vertical 3D, or a hybrid of both — fixes the achievable bandwidth and, just as importantly, the heat the rest of the system must remove. The next two sections take up that thermal and materials burden in turn.

SECTION THREE

3. Thermal Challenges & Heat Extraction

Every watt an accelerator consumes becomes heat that must leave through the package, and the harder the package works, the smaller and hotter the surface it must leave through. Thermal design is no longer a downstream packaging concern; it is a co-equal limit on performance.

The defining trend is rising **power density**. Accelerator power has climbed from tens of watts a decade ago toward many hundreds, and in some training parts beyond a kilowatt. Because advanced packaging concentrates compute into a small, dense footprint, the heat flux — watts per square centimetre — at the silicon surface has risen even faster than total power. Worse, the dissipation is rarely uniform: matrix-multiply units and high-traffic interconnect create local **hotspots** where flux is several times the die average, and it is the hotspot, not the average, that sets the throttling point.

Heat leaves the silicon along a **thermal resistance path** from the transistor junction to the ambient coolant. In a conventional package that path runs junction → silicon bulk → thermal interface material (TIM) → integrated heat spreader (lid) → a second TIM → the heat sink or cold plate → coolant. Each interface adds resistance, expressed in degrees Celsius per watt (°C/W); the junction temperature is the ambient plus the product of dissipated power and the summed resistance. Reducing any single dominant resistance lowers the whole stack's junction temperature, which is why the thermal interface materials and the cooling method receive disproportionate engineering attention.

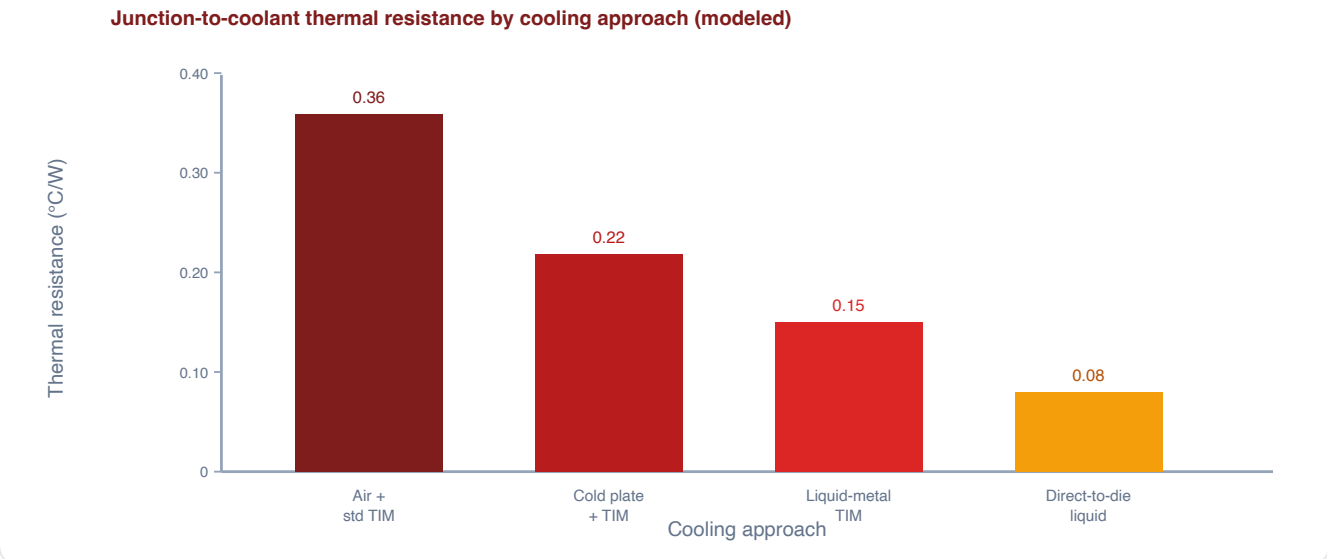

Figure 3.1 — Modeled junction-to-coolant thermal resistance for four cooling approaches. Lower is better; lid removal and direct liquid contact each strip a major resistance out of the stack. Illustrative values.

Several heat-extraction strategies follow from this picture. **Air cooling with a standard lid** is cheapest and simplest, but its resistance is dominated by the air-to-fin convection and by two TIM interfaces; it runs

out of headroom well below the power of a modern training accelerator. **Direct liquid cold plates** replace the air path with circulating coolant, sharply lowering the convective resistance. The next gains come from attacking the interfaces themselves: **direct-to-die (lidless) cooling** removes the integrated heat spreader and one TIM layer entirely, placing the cold plate against the bare silicon, while **near-junction cooling** goes further, etching microchannels or bonding microfluidic structures close to the heat-generating layer so the coolant works just micrometres from the hotspot.

> **DESIGN PRINCIPLE**
>
> Optimise for the hotspot, not the average. A package can be comfortable on its mean flux yet throttle because one matrix-multiply tile exceeds its junction limit. The cooling solution and the thermal interface material should be sized to the peak local heat flux and to the temperature-sensitive HBM, not to the die average.

For 3D-stacked parts the problem sharpens, because heat from a lower die must travel up through the dies above it before it can reach the spreader, and those upper dies are themselves heat sources. This is why memory-on-logic and logic-on-logic stacks place such a premium on thin dies, high-conductivity bonding, and, increasingly, on bringing the coolant inside the package. The material choices that govern every one of these resistances are the subject of Section 4.

SECTION FOUR

4. Materials: Interposers, Dielectrics & TIMs

Behind every packaging and thermal decision sits a materials trade-off. The interposer that carries the signals, the dielectric that insulates them, the spreader that conducts heat away, and the interface material that bridges the gaps each set hard limits — and each is a balance of performance against cost, manufacturability, and reliability.

Start with the **interposer and substrate**. A **silicon interposer** offers the finest wiring pitch and the closest thermal-expansion match to the dies it carries, which is why it dominates high-bandwidth-memory accelerators; its drawbacks are cost and a maximum size constrained by the same reticle limit that bounds logic dies. **Organic substrates** and **fan-out** approaches are far cheaper and can be larger, but support coarser wiring and expand differently from silicon under heat, stressing the joints between them. **Glass** interposers are an emerging middle path: dimensionally stable, capable of large panels, and electrically excellent, though still maturing in manufacturing. The choice ripples through the whole design, because it fixes both the achievable interconnect density and the thermomechanical stress the package must survive.

Within and around the wiring sit the **dielectrics**. As signals run faster and closer together, the insulating material between them must minimise capacitance and loss. **Low-k dielectrics** reduce that parasitic capacitance, lowering both delay and dynamic power, but lower-k materials are typically more porous and mechanically fragile, complicating the bonding and stacking that 3D integration requires. The dielectric choice is thus a quiet but constant negotiation between electrical performance and structural robustness.

The materials with the most direct thermal leverage are the **thermal interface materials (TIMs)** and the **heat spreaders**. A TIM fills the microscopic air gaps between two nominally flat surfaces — air being an excellent insulator — so heat can cross the boundary. Conventional polymer-and-filler thermal pastes are inexpensive but limited in conductivity. **Liquid-metal TIMs** (gallium-based alloys) conduct far better and can dramatically lower interface resistance, but they are electrically conductive and chemically aggressive toward aluminium, demanding careful containment. **Graphite** sheets spread heat anisotropically and are stable and clean to apply. For the spreader itself, copper is the workhorse, but **aluminium nitride (AlN)** and even synthetic **diamond** offer far higher conductivity for the most demanding near-junction applications — at a cost, and with their own integration challenges.

Thermal materials: indicative conductivity & trade-offs

Material	Thermal conductivity (W/m·K, indicative)	Role	Trade-off
Polymer TIM (paste)	~3 – 8	Interface fill	Cheap; limited conductivity, pump-out over time
Graphite sheet	~400 in-plane	Interface / spreading	Anisotropic; clean, stable
Liquid-metal TIM	~20 – 80	Interface fill	Excellent; conductive, corrosive, needs containment
Copper	~400	Spreader / lid	Standard; heavy, CTE mismatch with silicon
Aluminium nitride (AlN)	~150 – 200	Spreader / substrate	Good CTE match; brittle, costlier
Synthetic diamond	~1500 – 2000	Near-junction spreader	Highest conductivity; expensive, hard to integrate

No single material wins on every axis, which is the essence of the problem. Diamond conducts heat magnificently but is costly and difficult to bond; liquid metal is thermally superb but electrically hazardous; copper is reliable and cheap but expands differently from the silicon it cools, storing up the warpage stresses examined in Section 5. The art of the package is choosing a coherent set of materials whose strengths reinforce one another and whose weaknesses can be managed — a system optimisation, not a component-by-component one.

IN PRACTICE

The highest-conductivity material is rarely the right answer on its own. A modest TIM paired with a lidless, direct-liquid path often beats an exotic TIM trapped beneath a high-resistance lid. Materials selection only pays off when the entire resistance path — and the thermal-expansion compatibility of the whole stack — is optimised together.

SECTION FIVE

5. Yield & Reliability

A package that performs brilliantly but cannot be manufactured at economic yield is a laboratory curiosity. Yield and reliability are where advanced packaging earns its keep — and where disaggregation changes the arithmetic of what can profitably be built.

Yield begins with a statistical fact about defects. Manufacturing defects fall on a wafer at some average **defect density** (defects per square centimetre). The probability that a given die is free of killer defects falls as the die grows, because a larger die simply intercepts more of those defects. Under the widely used **Murphy yield model**, yield declines steeply and non-linearly with die area for any fixed defect density. The practical consequence is severe: doubling a die's area does far more than halve its yield, which is why very large monolithic dies become disproportionately expensive — most of the wafer is scrapped.

Disaggregation into chiplets directly attacks this. By splitting a large logic block into several smaller dies, each chiplet individually enjoys the much higher yield of a small die. Defective chiplets are discarded cheaply before assembly, and only good dies are integrated, so the yield loss is confined to small pieces rather than to an entire expensive monolith. This is the single most powerful economic argument for chiplets, and it is quantified in the model of Section 6.

That advantage is not free. It depends on **known-good-die (KGD) testing** — the ability to test each chiplet thoroughly *before* it is committed to an expensive package, because once dies are bonded together, a single bad chiplet can scrap the whole assembly, along with the good dies and costly interposer beside it. KGD testing of bare dies, and especially of dies destined for 3D stacks where they cannot be reworked, is one of the harder problems in the field.

Assembly introduces its own yield terms. The package now depends on tens of thousands of **microbump and hybrid-bond interconnections**, each of which must form correctly; **interconnect yield** therefore becomes a meaningful factor in its own right, and it worsens as bump pitch shrinks and bump count rises. The more chiplets and the more interconnects, the more places assembly can fail — a counterweight to the die-yield gains of disaggregation that the overall model must balance.

Finally there is **thermomechanical reliability**. Different materials in the stack — silicon, copper, the organic substrate, the underfill — expand by different amounts when heated, a mismatch in coefficient of thermal expansion (CTE). As the package heats and cools through manufacturing and through every power cycle in the field, those mismatches induce stress that bends the package (**warpage**) and fatigues the solder joints and bumps. Excessive warpage can crack low-k dielectrics, open interconnects, or prevent the package from seating flat against its cold plate, undoing the thermal design. Managing warpage — through underfill selection, stiffeners, balanced stack-ups, and CTE-matched materials — is a central reliability task, and it grows harder as dies thin and stacks rise.

THE YIELD EQUATION, RESTATED

Disaggregation trades one large, low-yielding die for several small, high-yielding ones plus a new assembly-yield and reliability burden. The net benefit is real and often large, but it holds only when known-good-die testing is rigorous and the interconnect and warpage budgets are respected. Chiplets do not abolish yield risk; they relocate it from the wafer to the package.

SECTION SIX

6. Methodology: Thermal & Yield Model

The figures in this whitepaper come from two simple, transparent models — one thermal, one for yield — built from textbook relationships and representative parameters. The models are illustrative tools for reasoning about trade-offs, not predictions for any specific product. Every assumption is stated so the reader can substitute their own.

The **thermal model** treats the package as a one-dimensional resistance network. The steady-state junction temperature is the coolant temperature plus the dissipated power multiplied by the summed junction-to-coolant thermal resistance. We decompose that resistance into its dominant terms — silicon bulk, the thermal interface material(s), the spreader or lid, and the convective or liquid path to coolant — and vary the cooling approach to see its effect on junction temperature. A hotspot multiplier captures the fact that local flux can exceed the die average.

```
# 1-D junction temperature model (steady state)
T_junction = T_coolant + power_W * R_total          # degrees C
R_total    = R_silicon + R_TIM1 + R_lid + R_TIM2 + R_coolant
T_hotspot  = T_coolant + (power_W * hotspot_factor) * R_total
# throttle when T_hotspot exceeds T_max (e.g. 105 C); HBM limit is lower
def headroom(power_W, R_total, T_coolant, T_max):
    return T_max - (T_coolant + power_W * R_total)
```

Listing 6.1 — The thermal model. Each resistance term is an input; lidless and direct-liquid paths zero out R_lid and one TIM layer.

The **yield model** uses the Murphy approximation, in which die yield is a function of die area and defect density. The yield of a multi-chiplet design is the product of the chiplet yields (each computed at its own small area) and an assembly-and-interconnect yield term that degrades with the number of interconnections. This lets us compare a single large monolithic die against an equivalent design split into several chiplets, on equal defect-density assumptions.

```
# Murphy die-yield model + chiplet composition
def murphy_yield(area_cm2, defect_density):
    x = area_cm2 * defect_density
    return ((1 - exp(-x)) / x) ** 2          # Murphy approximation

monolithic_Y = murphy_yield(A_total, D0)
chiplet_Y    = murphy_yield(A_total / n, D0) ** n * assembly_yield(n_bumps)
# compare effective good-die cost per unit compute
```

Listing 6.2 — The yield model. n chiplets of area A_total/n each, composed with an assembly-yield penalty that rises with interconnect count.

Modeled input assumptions (illustrative defaults)

Parameter	Value	Parameter	Value
Accelerator power	700 W	Coolant temperature	40 °C
Junction limit (logic)	105 °C	Hotspot factor	2.5× die average
Defect density (D_0)	0.10 / cm²	Reference die area	up to 800 mm²
Chiplet split	1 vs 4 dies	Assembly yield (per interconnect)	illustrative, >99.99%

These figures are representative of a high-power training-class accelerator and are *illustrative, not field-measured*. The purpose is a transparent, adjustable model — not a claim about any specific device or process. The charts in Section 7 are generated from these two models on the assumptions above.

SECTION SEVEN

7. Findings

Three results recur across the modeled scenarios: thermal resistance, not transistor count, sets the sustained-performance ceiling; yield collapses with die area unless the design is disaggregated; and the two effects together favour chiplet-based packages with aggressive cooling.

~78%
JUNCTION-TO-COOLANT R CUT, AIR→DIRECT-DIE

800 mm²
RETICLE-LIMITED MONOLITHIC CEILING

~2×
GOOD-DIE YIELD GAIN VIA 4-CHIPLET SPLIT

2.5×
HOTSPOT FLUX VS DIE AVERAGE

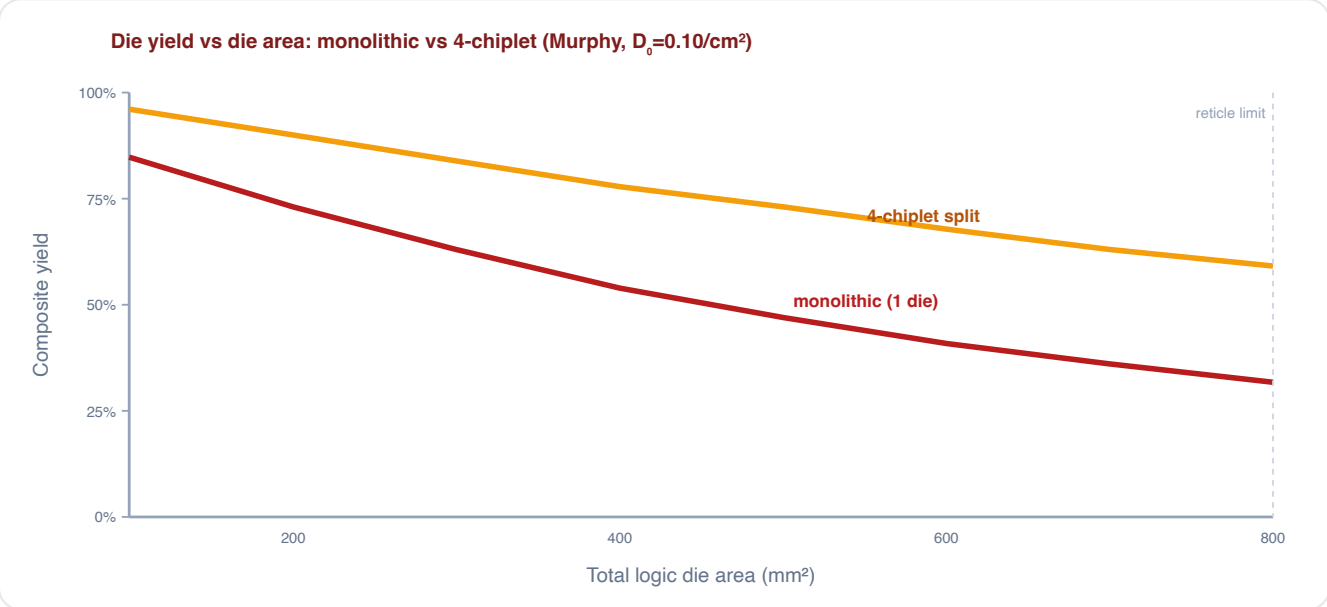

Figure 7.1 — Modeled composite yield vs total logic area. Splitting one large die into four chiplets recovers a substantial share of the yield that monolithic scaling surrenders. Illustrative; assembly-yield penalty included.

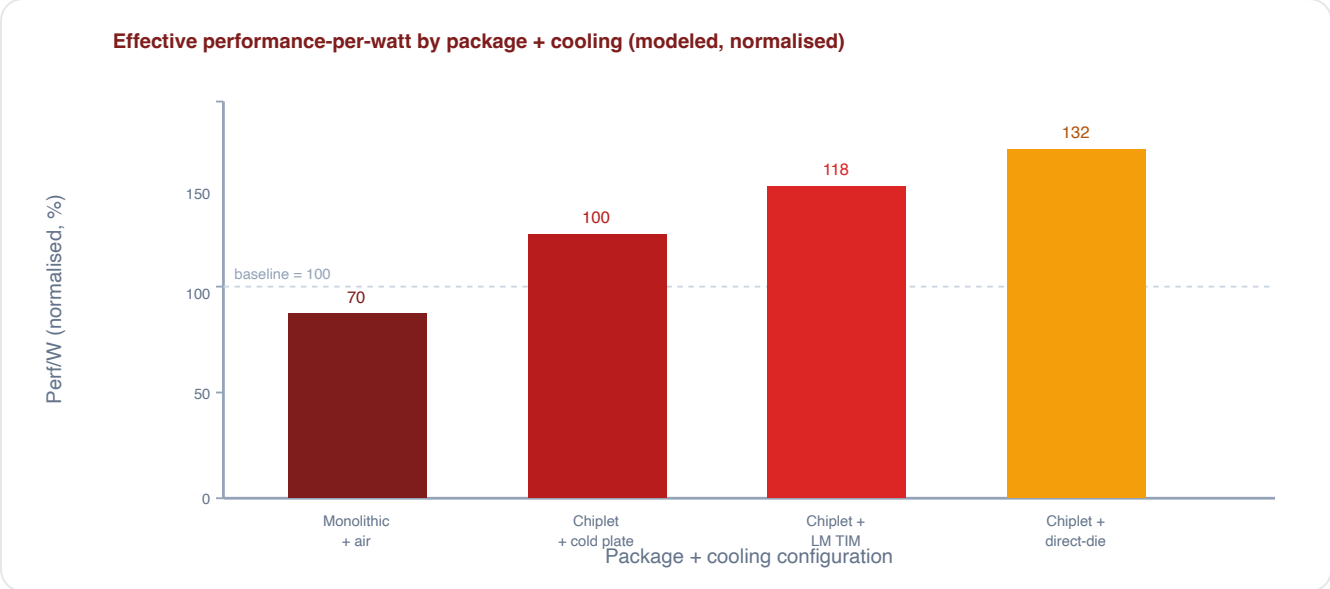

Figure 7.2 — Modeled effective performance-per-watt, normalised to a chiplet + cold-plate baseline. Lower thermal resistance lets the part sustain its clocks instead of throttling, raising delivered perf/W. Illustrative values.

The yield curve (Figure 7.1) is the central economic finding. A single large monolithic die approaching the reticle limit yields only about a third of good parts under the modeled defect density, whereas the same total logic area split into four chiplets yields close to twice as many usable devices, even after charging an assembly-yield penalty for the additional interconnections. This is the quantitative heart of the chiplet argument: disaggregation converts a steep, punishing area-yield curve into a much gentler one.

The thermal-resistance result (Figure 3.1) and the performance-per-watt comparison (Figure 7.2) tell the complementary story. Moving from air cooling with a standard lid to a lidless, direct-to-die liquid path cuts modeled junction-to-coolant resistance by roughly three-quarters. Because junction temperature scales directly with that resistance, the same silicon at the same power runs far cooler — and a cooler part can hold its clocks rather than throttling, which is why the delivered performance-per-watt rises sharply with each step down the resistance ladder. The two effects compound: a chiplet design both yields better and, with aggressive cooling, sustains higher performance.

RESULT SUMMARY

- Composite yield for a reticle-limit monolithic die (~32% modeled) roughly doubles when the design is split into four chiplets, net of assembly penalties.
- Junction-to-coolant thermal resistance falls ~78% from air-plus-lid to direct-to-die liquid, lowering junction temperature at fixed power proportionally.
- Sustained performance-per-watt rises with each reduction in thermal resistance because the part throttles less; the hotspot, at ~2.5× die average, governs the throttle point.
- The chiplet and thermal gains compound: disaggregated, aggressively cooled packages dominate the modeled buildability-and-efficiency frontier.

SECTION EIGHT

8. Discussion, Limitations & Outlook

A model is a lens, not a measurement. The results above should be read as structure — the shape of the curves and the rank-order of the levers — rather than as precise predictions for any one product or process.

Limitations. The thermal model is one-dimensional and steady-state; it captures the dominant resistance path and a hotspot multiplier but not transient thermal behaviour, lateral spreading, or the detailed three-dimensional heat flow inside a stacked package, all of which a real design would simulate with finite-element methods. The yield model uses the Murphy approximation with a single fixed defect density and an illustrative assembly-yield term; real fabs see defect densities that vary by node maturity and defect type, and assembly yield depends intricately on bump pitch, bonding method, and test coverage. The performance-per-watt figures are normalised and qualitative, meant to show direction, not to rank named products. We have also held many second-order effects fixed — redundancy and repair, binning, partial-good harvesting, and the design and verification overhead of multi-chiplet systems — several of which would further shift the economics, mostly in favour of disaggregation.

Key risks and challenges. The dominant challenges in advanced packaging are as much manufacturing and supply-chain as physics. Advanced-packaging capacity, particularly silicon-interposer platforms, is a real bottleneck and a strategic constraint on the whole accelerator industry. Known-good-die testing of bare and stacked dies remains hard, and a single bad chiplet can scrap an expensive assembly. Warpage and thermomechanical fatigue grow more dangerous as dies thin and stacks rise. And the thermal envelope is becoming a data-centre-level problem: parts that demand direct liquid or near-junction cooling force change far upstream into facility design.

Outlook. The trajectory is clear. Interconnect will continue to densify, with hybrid bonding pushing toward ever-finer pitch and true 3D logic-on-logic stacks. Standard die-to-die interfaces such as UCIe will deepen the chiplet ecosystem, letting designers mix dies from multiple sources and nodes. Cooling will move inside the package, with microfluidics and near-junction structures becoming mainstream for the highest-power parts. Glass interposers and new spreader materials will widen the materials palette. Through all of it, the package — not the transistor alone — will remain where accelerator performance is won or lost.

IN PRACTICE

The most reliable predictor of a successful accelerator program is not the process node chosen but whether thermal and yield were treated as first-class design inputs from the outset — co-designed with the architecture — rather than as packaging problems handed downstream after the floorplan was frozen.

Recommendations

1. **Treat the package as architecture.** Decide 2.5D versus 3D, and the chiplet partitioning, alongside the logic floorplan — not after it.
2. **Disaggregate large logic into chiplets** wherever yield economics and the reticle limit bite, but budget honestly for known-good-die testing and assembly yield.
3. **Design for the hotspot, not the average,** and choose the cooling path — up to lidless direct-to-die or near-junction liquid — to the peak local flux and the temperature-sensitive HBM.
4. **Optimise the whole resistance path together;** an exotic TIM trapped under a high-resistance lid wastes its potential. Match thermal-expansion behaviour across the stack to control warpage.
5. **Co-design with the data centre.** If the part needs liquid cooling, that requirement reaches into rack and facility design and must be planned, not discovered.

Conclusion

The transistor is no longer the sole engine of accelerator performance. As monolithic scaling slows, the package has become the frontier — and on that frontier, two constraints dominate: how much heat can be pulled out of an increasingly dense and powerful die, and how many good devices a wafer and assembly line can economically produce. Advanced packaging answers both, escaping the reticle limit, recovering yield through chiplet disaggregation, and placing memory beside logic for the bandwidth AI demands. But it does so by relocating difficulty into thermal extraction, materials, interconnect, and warpage. Teams that treat thermal and yield as first-class, co-designed inputs — choosing geometry, materials, and cooling as one system — will build accelerators that are both manufacturable and fast. Those who leave packaging to the end will find that the heat, or the yield, has quietly capped the very performance they set out to win.

References

1. Moore, G. E. — Cramming more components onto integrated circuits (1965); and subsequent revisions on density scaling.
2. Dennard, R. H. et al. — Design of ion-implanted MOSFETs with very small physical dimensions (1974).
3. Murphy, B. T. — Cost-size optima of monolithic integrated circuits (yield modelling).
4. JEDEC — High Bandwidth Memory (HBM) DRAM standard (JESD235 series).
5. UCIe Consortium — Universal Chiplet Interconnect Express (UCIe) Specification.
6. Root Digit Advanced Hardware Lab — Internal thermal-and-yield modelling methodology (2026).

-rootdigit®

Performance lives in the package now.

Root Digit designs and evaluates advanced-packaging and thermal systems for AI accelerators — 2.5D and 3D integration, chiplet partitioning, thermal-resistance and yield modelling, and materials selection from interposer to cold plate. We bring the model; you keep the headroom.

Talk to our hardware engineers →

rootdigit.com

General: info@rootdigit.com · **Consultation:** rootdigit.com/contact/consultation
Dubai International Financial Centre (DIFC) · Wyoming, USA