



DESIGNING BEYOND THE RETICLE

WAFER-SCALE ARCHITECTURE AND THE FUTURE OF
HETEROGENEOUS INTEGRATION

Jean-Philippe Fricker

Co-Founder & Chief System Architect

Cerebras Systems

LEGAL DISCLAIMER

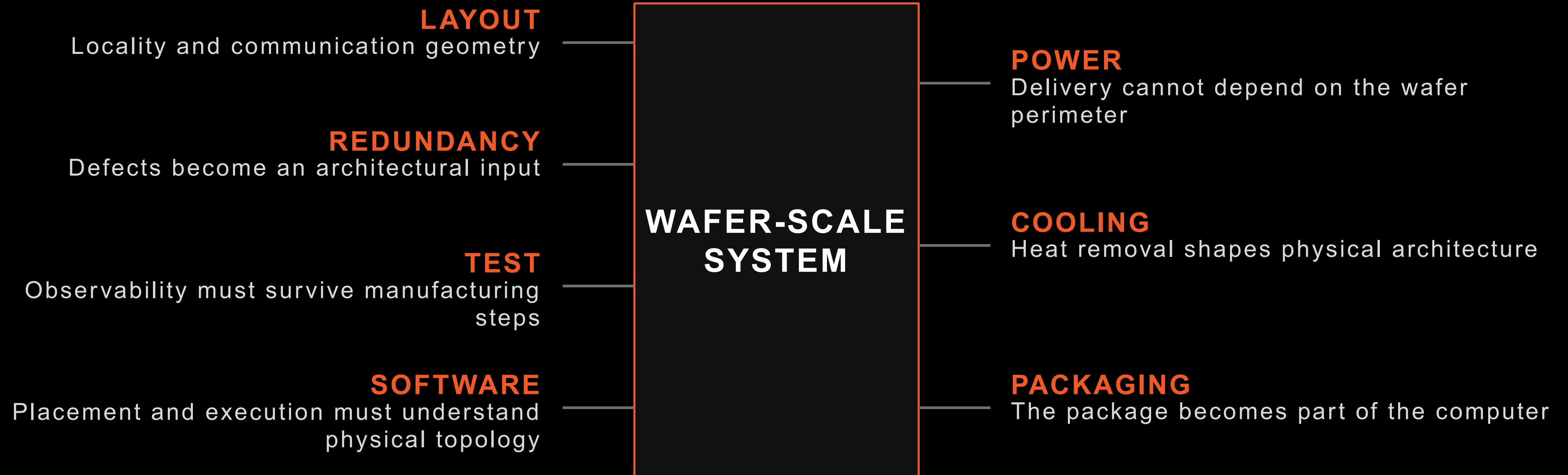
Forward-Looking Statements

This presentation and the accompanying oral presentation contain “forward-looking statements” within the meaning of applicable securities laws. All statements other than statements of historical fact could be deemed to be forward-looking. The words “may,” “will,” “shall,” “should,” “expects,” “plans,” “anticipates,” “could,” “intends,” “target,” “projects,” “contemplates,” “believes,” “estimates,” “predicts,” “potential,” “objective,” or “continue,” or the negative of these words or other similar terms or expressions that concern our expectations, strategy, plans, or intentions are intended to identify forward-looking statements, although not all forward-looking statements contain these identifying words. These forward-looking statements are subject to a number of risks and uncertainties, many of which involve factors or circumstances that are beyond Cerebras’ control.

Cerebras’ actual results could differ materially from those stated or implied in forward-looking statements due to a number of factors. Accordingly, undue reliance should not be placed on such statements. These forward-looking statements are made as of the date of this presentation and are based on information available to Cerebras together with Cerebras’ expectations, estimates, forecasts, projections, beliefs, and assumptions as of such date. These forward-looking statements should not be relied upon as representing Cerebras’ views as of any date subsequent to the date of this presentation and the accompanying oral presentation. Past performance is not necessarily indicative of future results. Cerebras undertakes no intention or obligation to update or revise any forward-looking statements, whether as a result of new information, future events, or otherwise, except as required by law.

Further information on potential risks that could affect actual results is included in Cerebras’ most recent filings with the Securities and Exchange Commission (the “SEC”), including in Cerebras’ most recent Quarterly Report on Form 10-Q, copies of which may be obtained by visiting Cerebras’ Investor Relations website at investors.cerebras.ai or the SEC’s website at www.sec.gov.

BEYOND THE RETICLE — EVERY DESIGN DOMAIN BECOMES COUPLED



WAFER SCALE IS NOT A LARGER CHIP — IT IS A DIFFERENT SYSTEM-DESIGN DISCIPLINE

WHY EVEN BOTHER THEN?

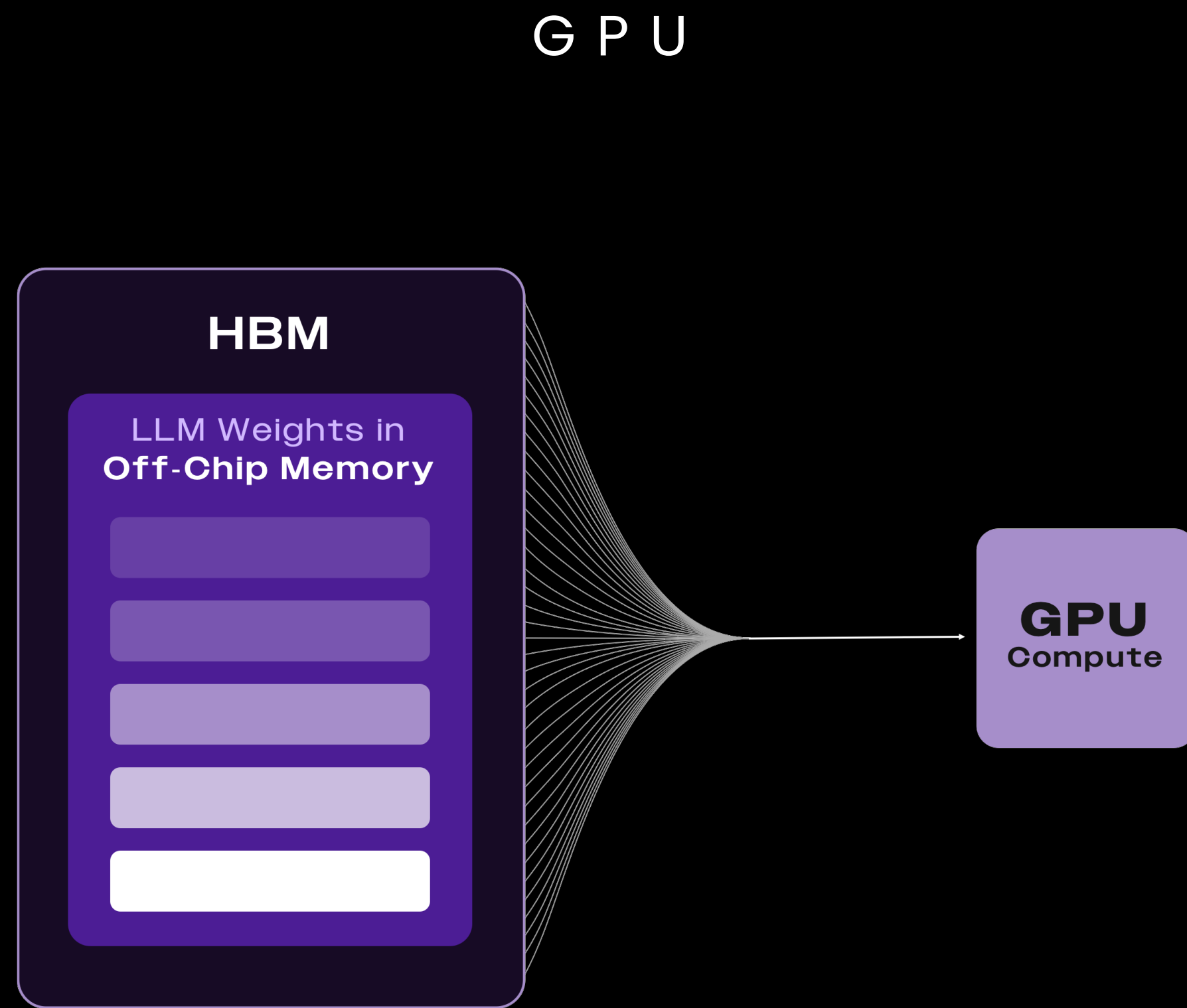
THE PROBLEM THAT BECAME CEREBRAS

1,000×
more AI compute

We did not begin by asking how to build a larger chip.

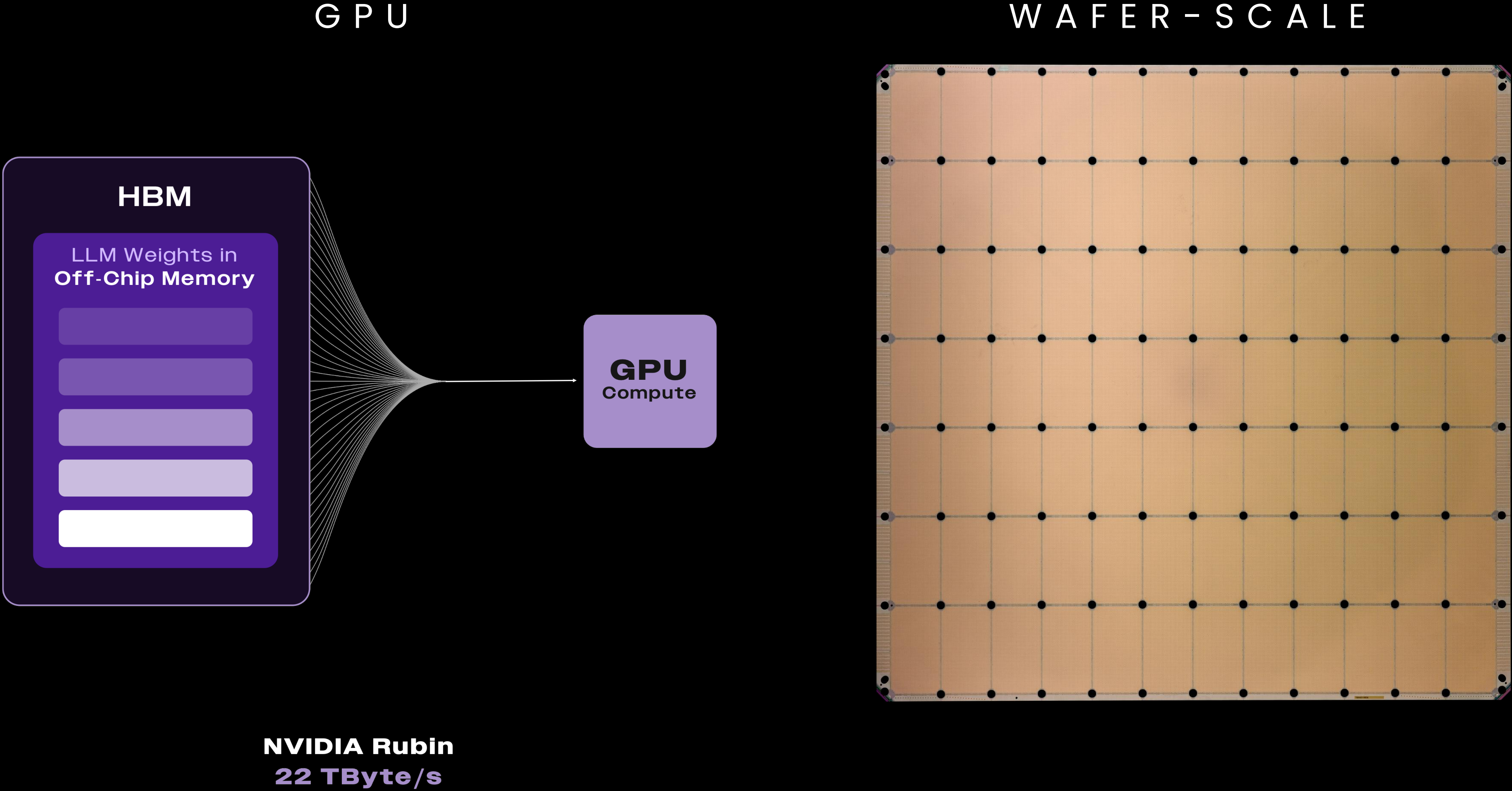
We began with the performance gap
— and discovered that wafer scale was the answer.

WAFER SCALE SHIFTS THE PERFORMANCE BOUNDARY FROM LOCAL MEMORY BANDWIDTH TO SYSTEM I/O

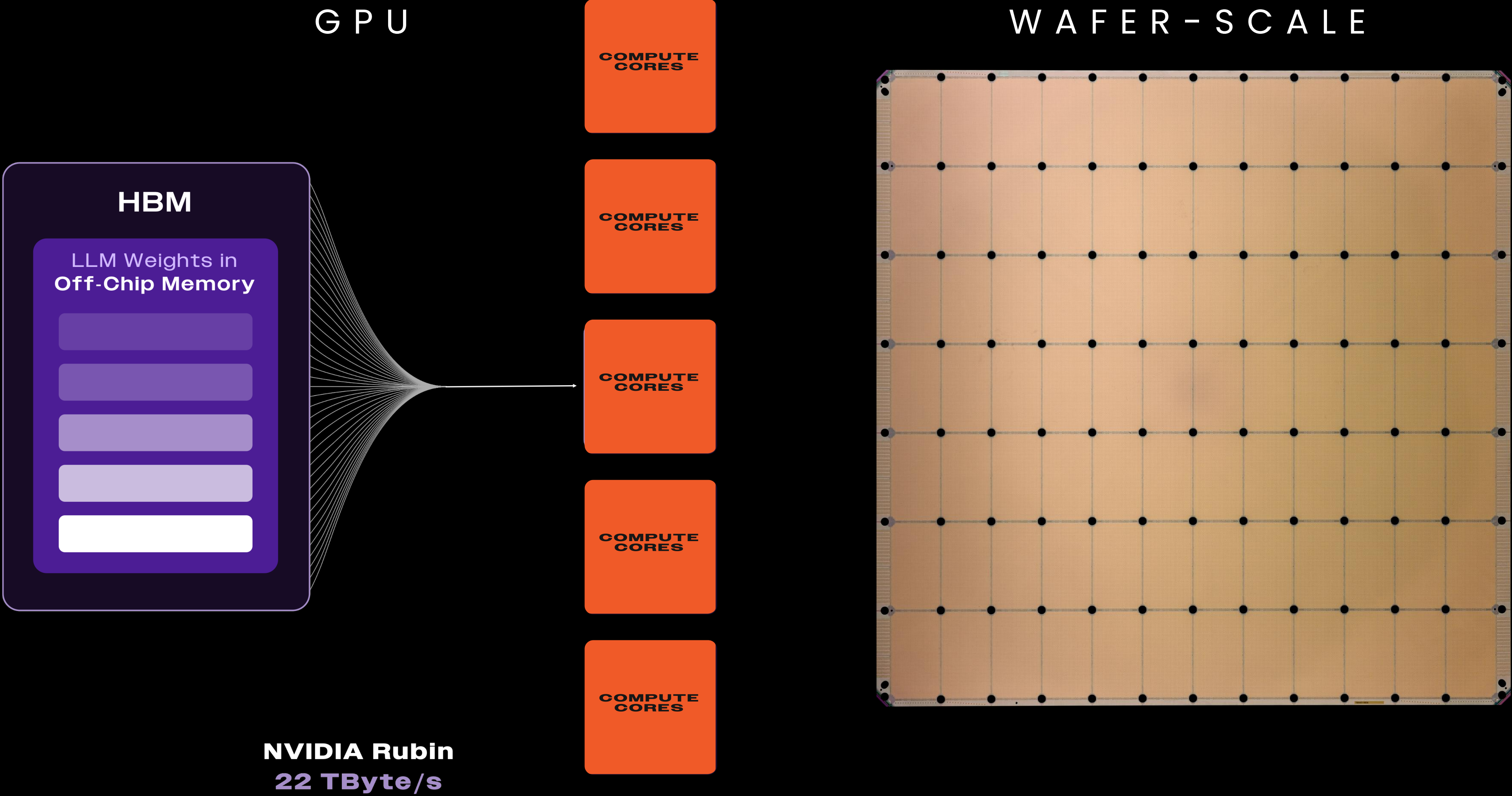


NVIDIA Rubin
22 TByte/s

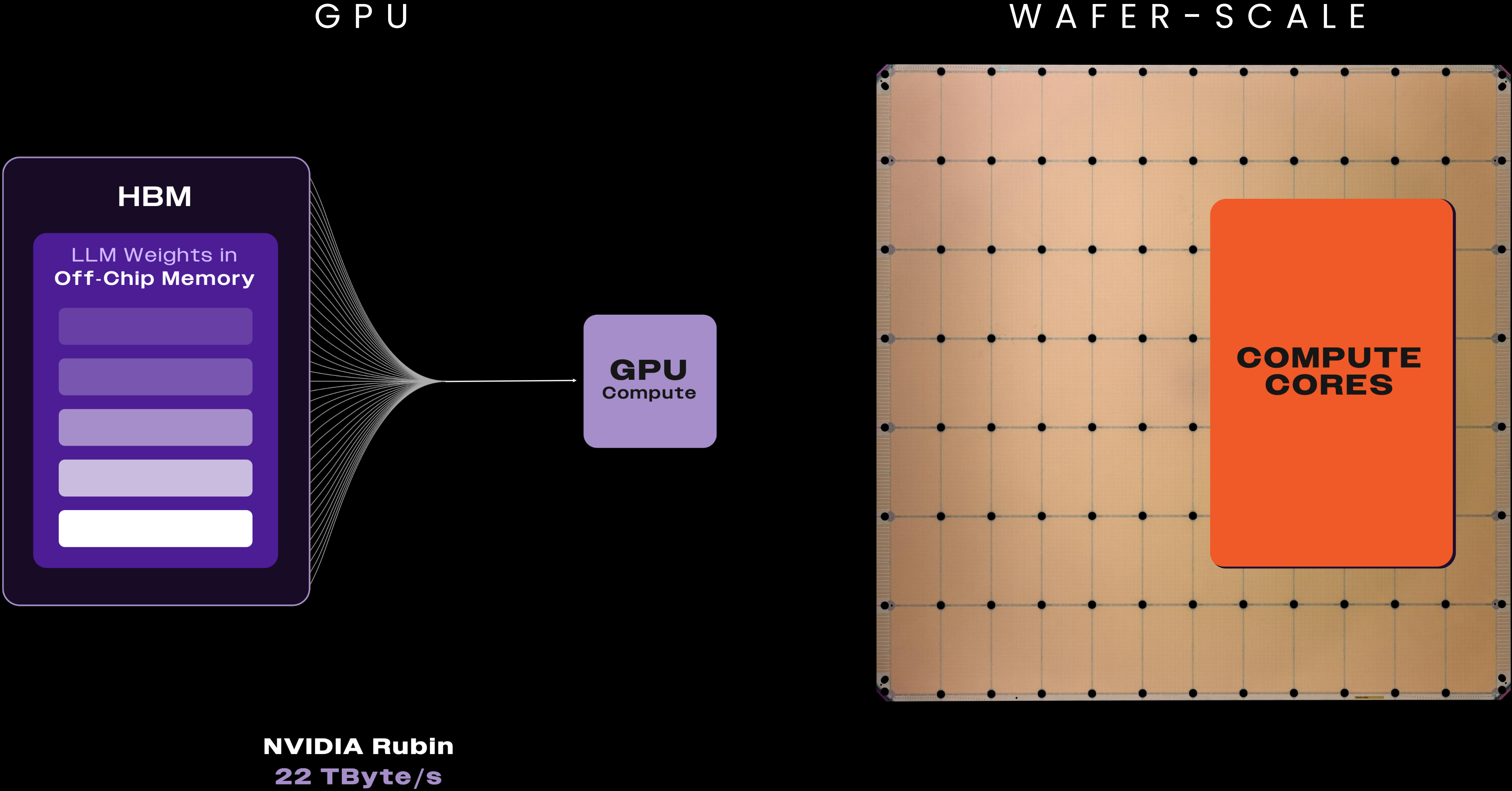
WAFER SCALE SHIFTS THE PERFORMANCE BOUNDARY FROM LOCAL MEMORY BANDWIDTH TO SYSTEM I/O



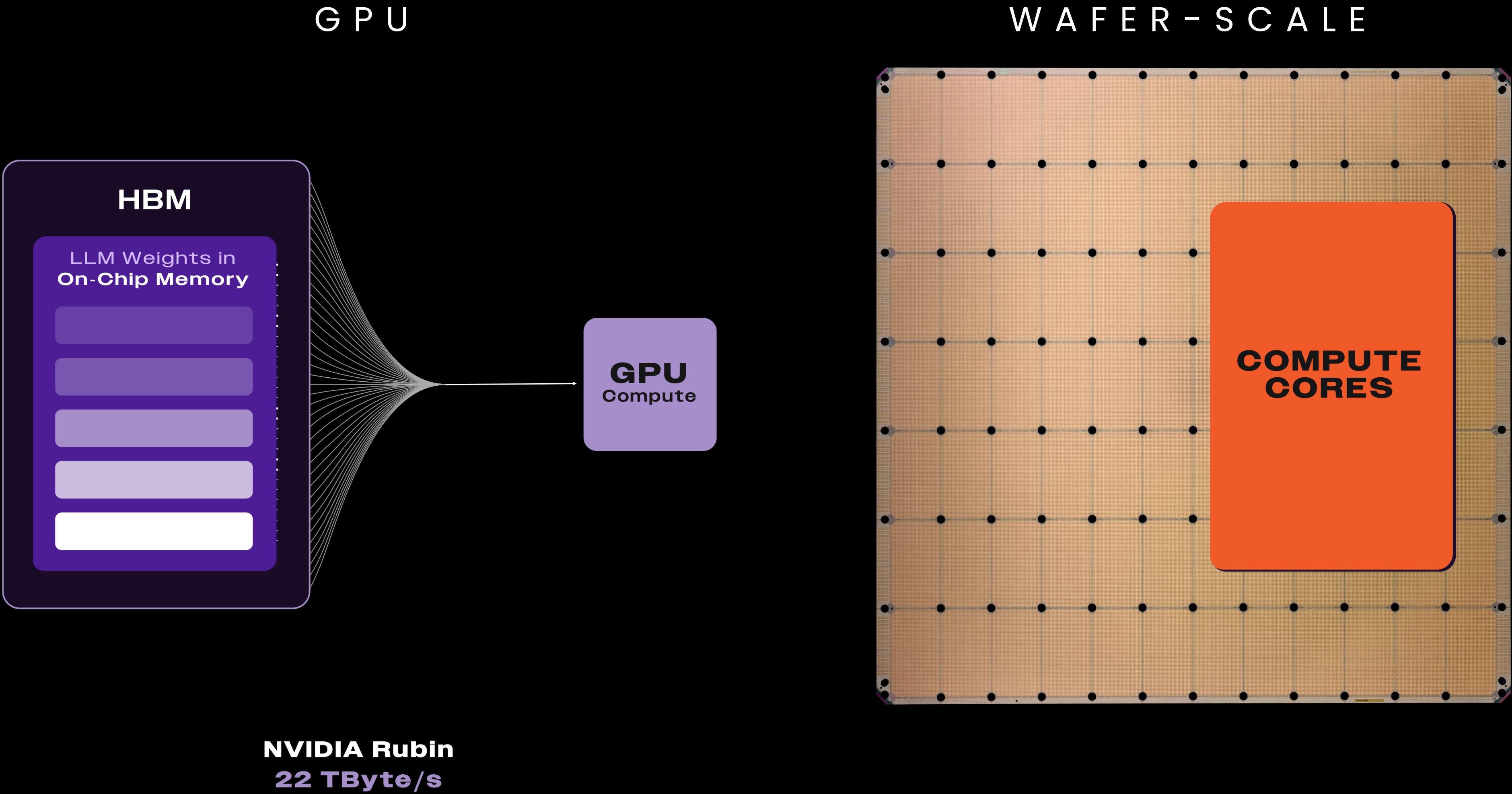
WAFER SCALE SHIFTS THE PERFORMANCE BOUNDARY FROM LOCAL MEMORY BANDWIDTH TO SYSTEM I/O



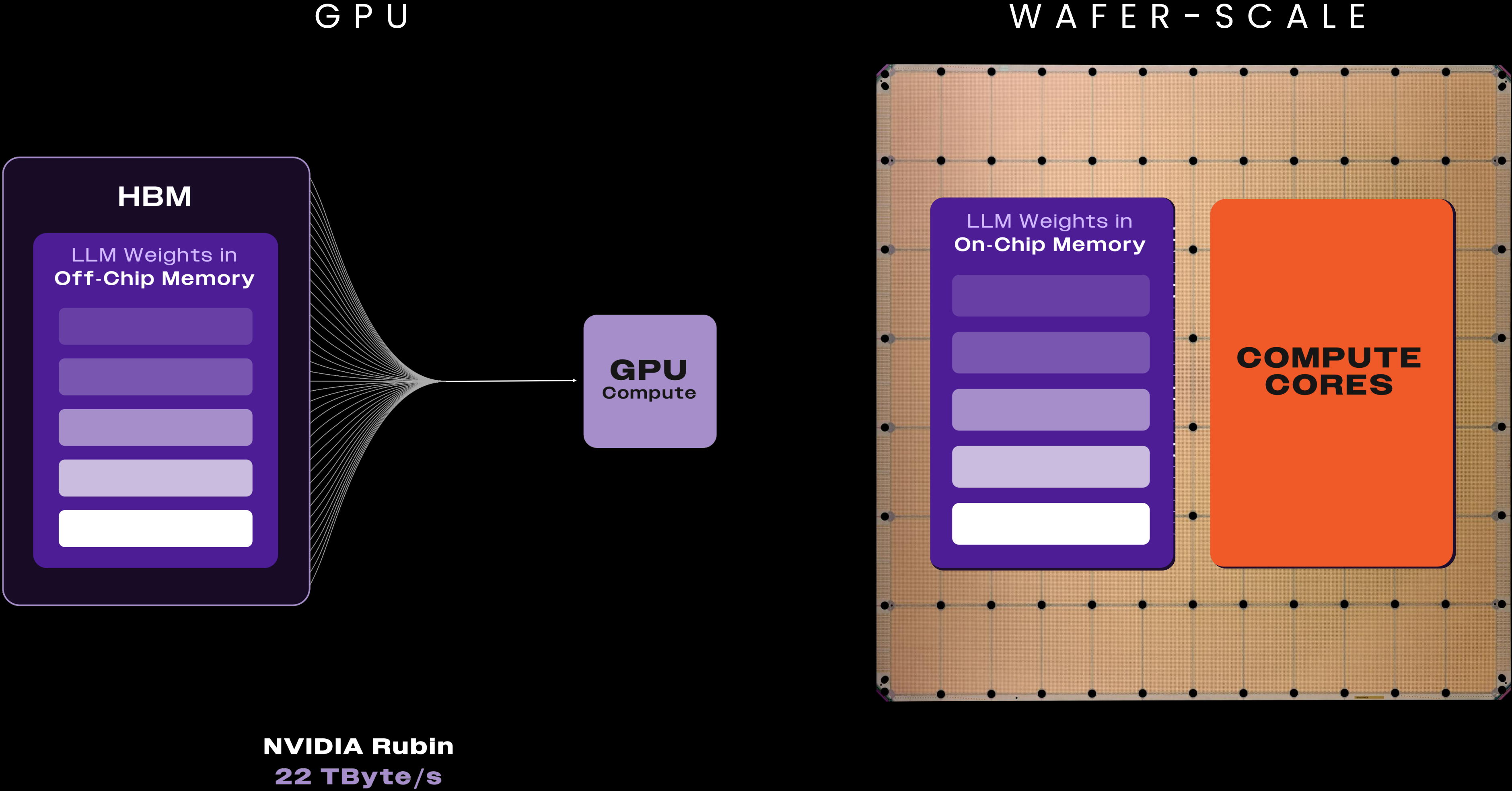
WAFER SCALE SHIFTS THE PERFORMANCE BOUNDARY FROM LOCAL MEMORY BANDWIDTH TO SYSTEM I/O



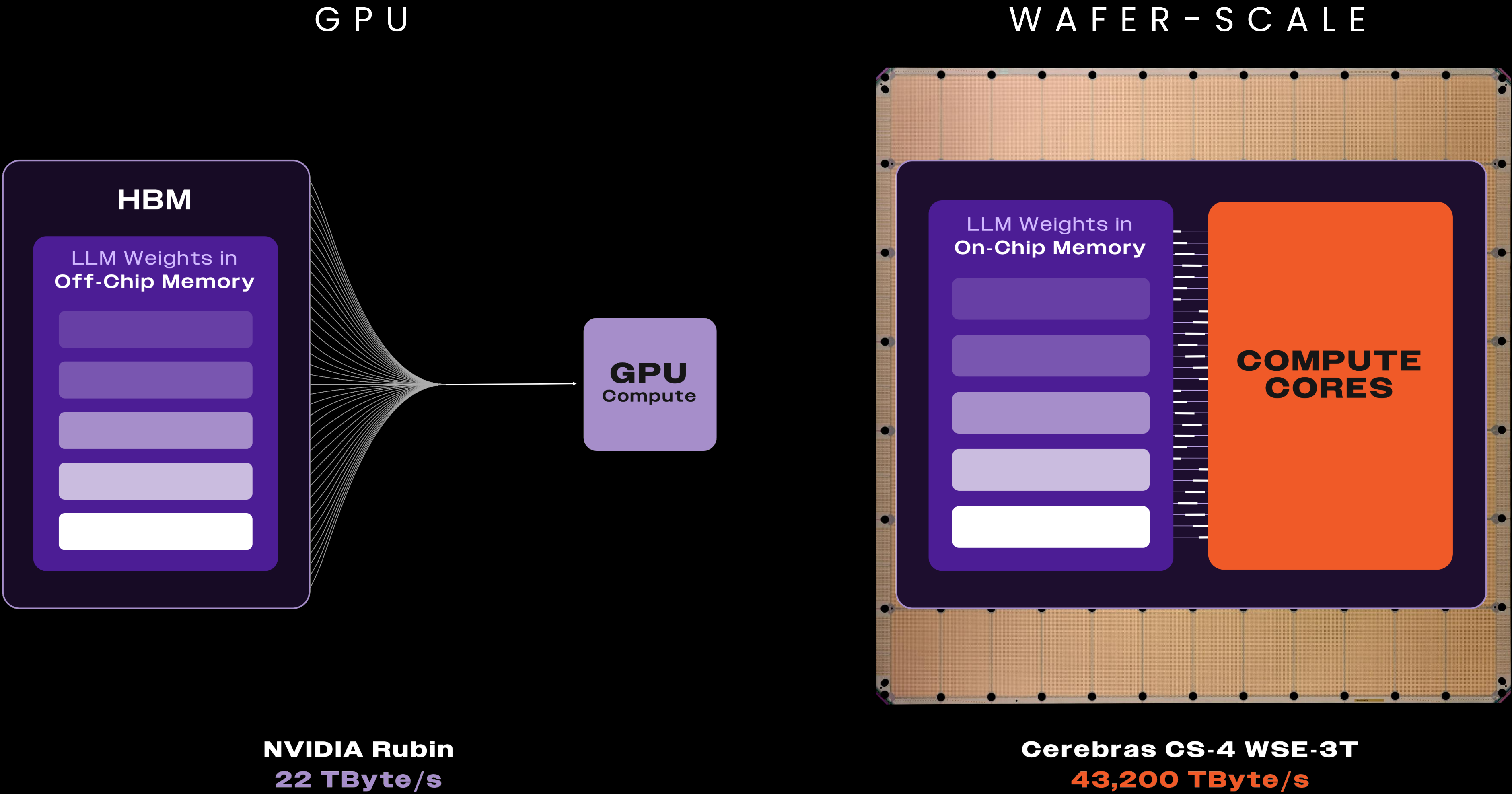
WAFER SCALE SHIFTS THE PERFORMANCE BOUNDARY FROM LOCAL MEMORY BANDWIDTH TO SYSTEM I/O



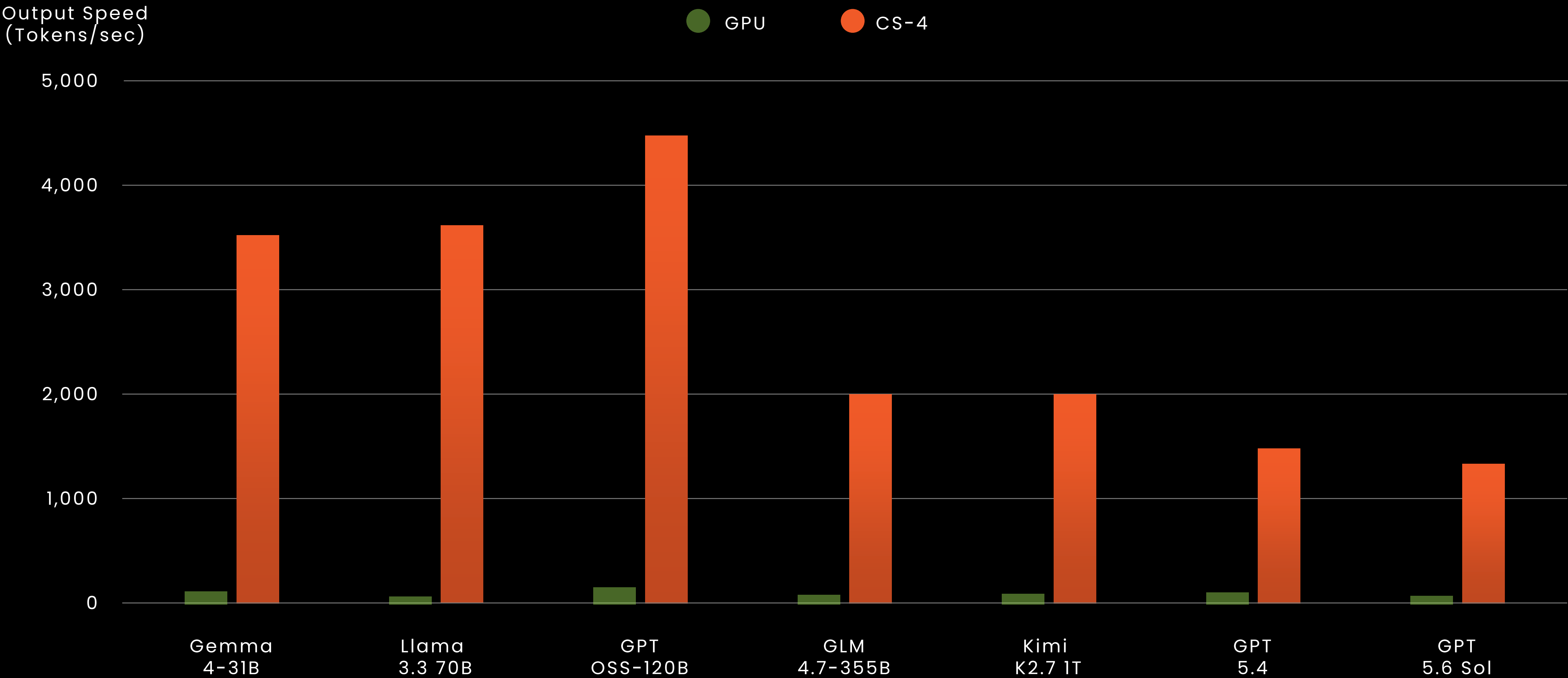
WAFER SCALE SHIFTS THE PERFORMANCE BOUNDARY FROM LOCAL MEMORY BANDWIDTH TO SYSTEM I/O



WAFER SCALE SHIFTS THE PERFORMANCE BOUNDARY FROM LOCAL MEMORY BANDWIDTH TO SYSTEM I/O

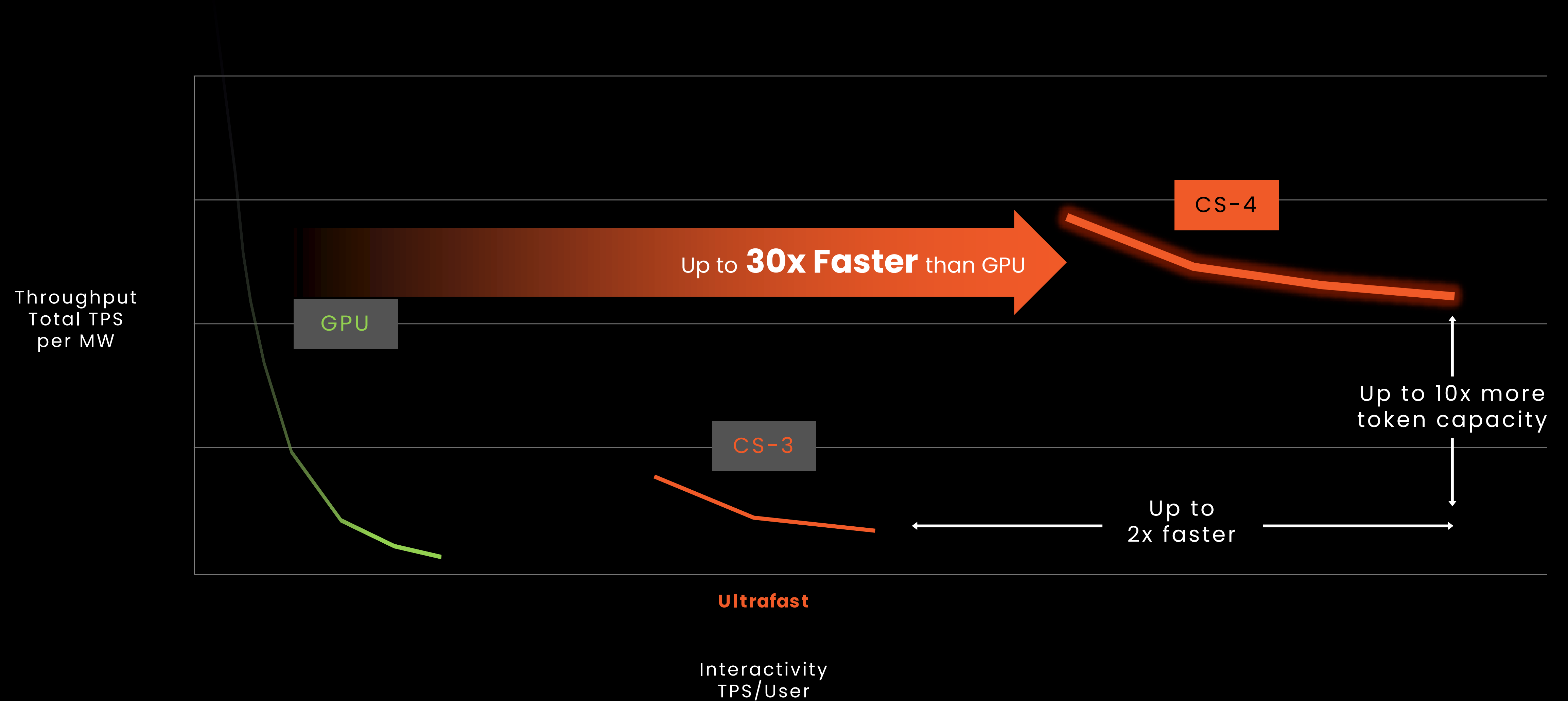


CS-4 IS TRANSFORMATIVE: *UP TO 30X FASTER*



Source: Artificial Analysis and internal benchmarking (August 2026)

THE FASTEST GETS FASTER. BUILT FOR HYPERSCALE.



Source: Artificial Analysis and internal benchmarking (August 2026)

CS-4 Speed Advantage



THE PROBLEM THAT BECAME CEREBRAS

1,000×
more AI compute

We did not begin by asking investors to finance a larger chip.

We began with the performance gap — then showed why wafer scale was the answer.

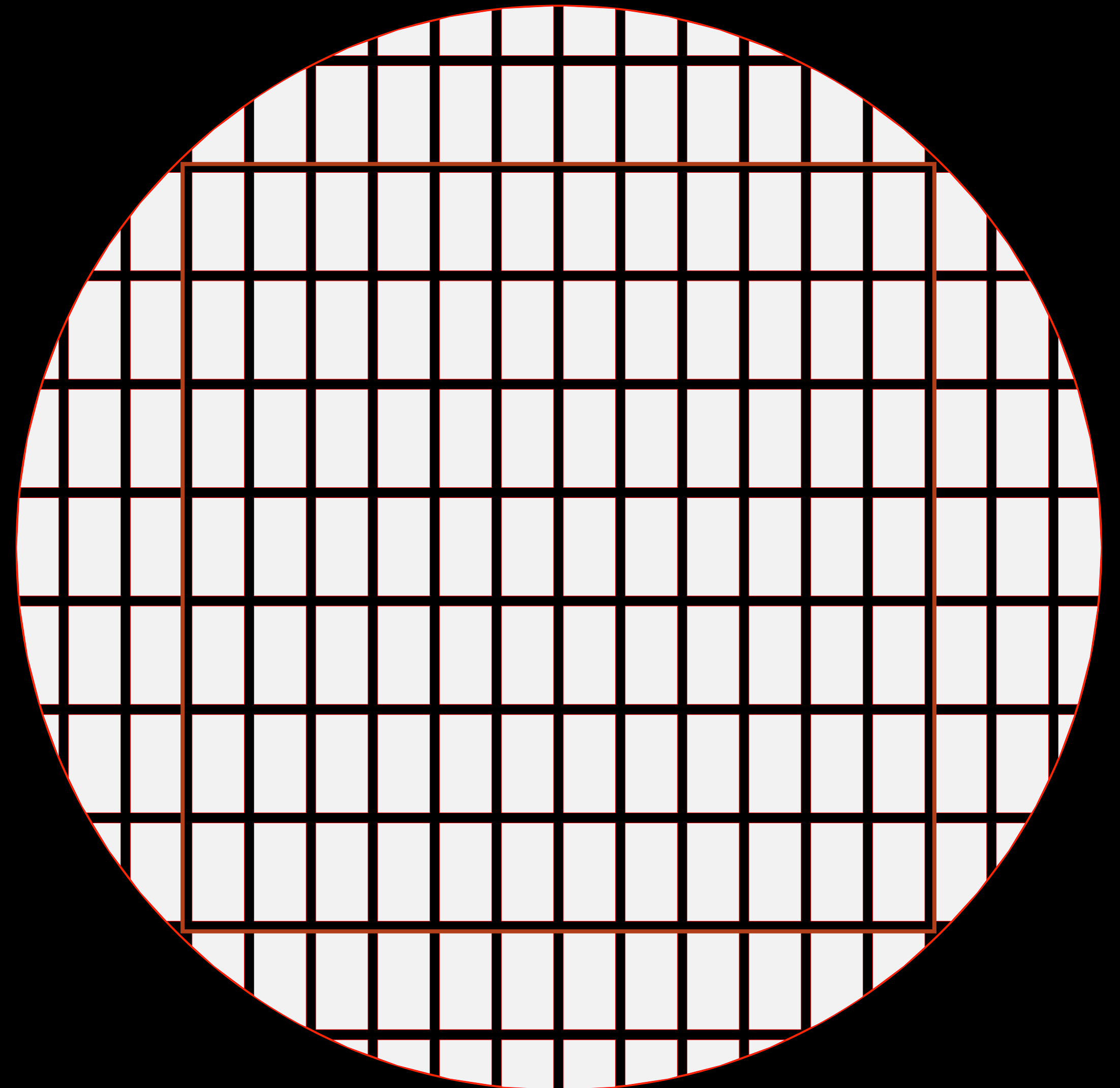
**Lead with why.
Earn the right
to explain how.**

THE UNMUTABLE MASK SET WALL

- Standard fabrication process requires the same mask set to be stepped and repeated across the entire wafer

SAME MASK SET

- To reduce variations of metal densities
- To reduce time at stepper to swap mask sets



THE UNPENETRABLE SCRIBE LINE AND SEAL RING WALLS

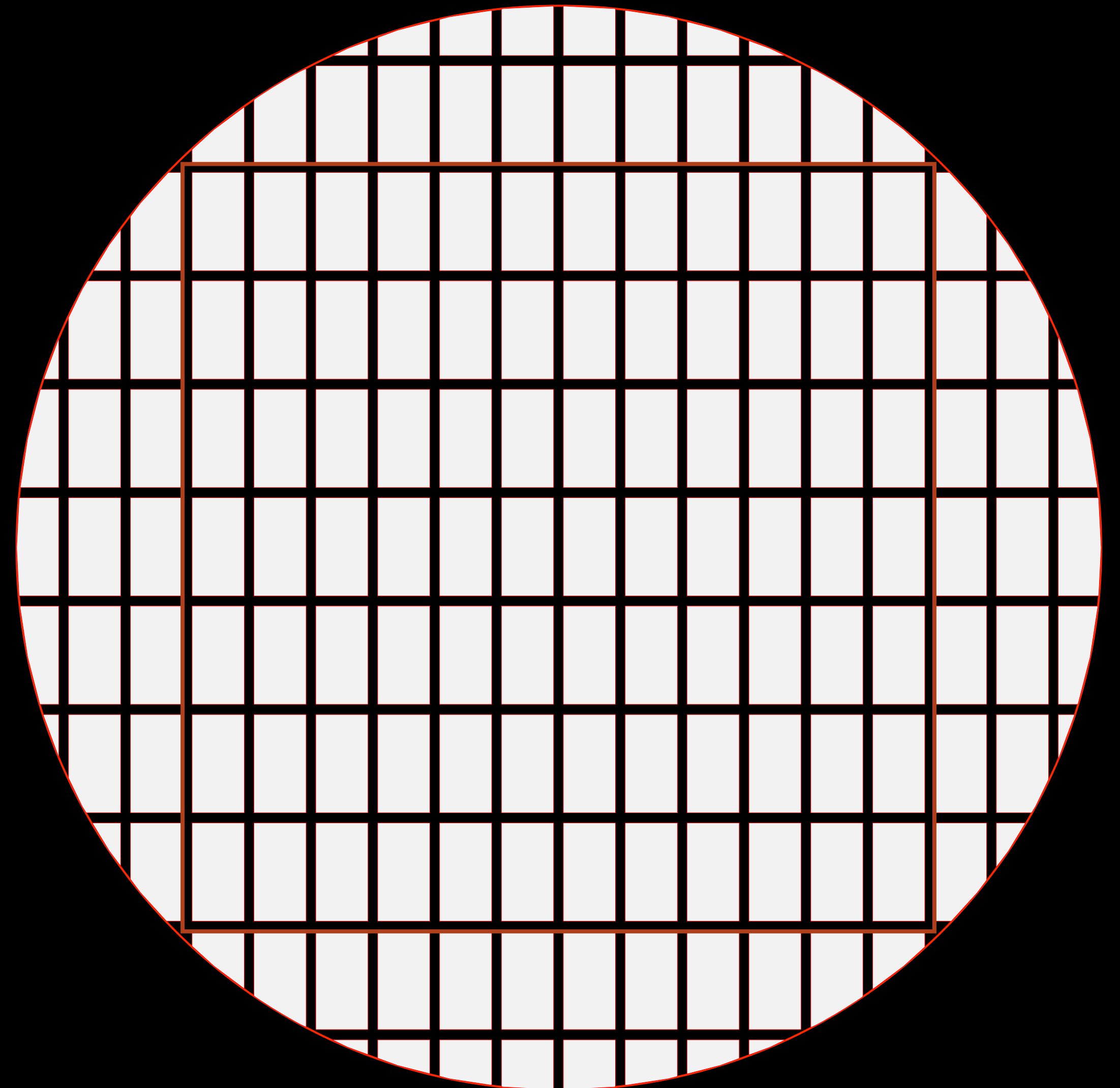
- Standard fabrication process requires die to be independent

SCRIBE LINES

- Scribe line separates each die
- Scribe line used as mechanical barrier for die cutting and for FAB test structures

SEAL RINGS

- Seal ring surrounds each reticles
- Metal barriers on each BEOL layer
- Essential at cut die level

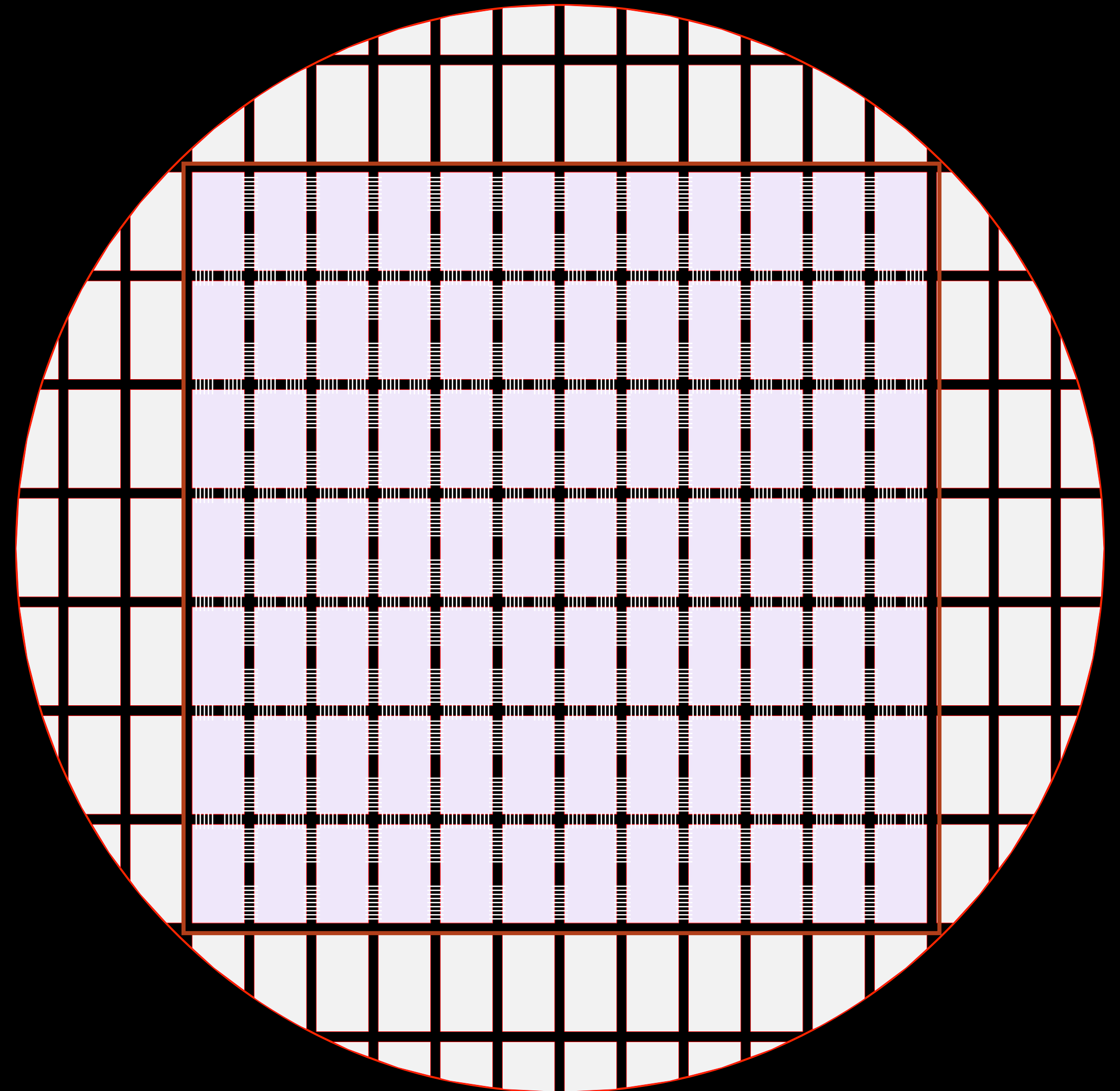


BEYOND THE RETICLE WITH CROSS SCRIBE LINE WIRES

CROSS SCRIBE LINE WIRES

- Add wires across scribe line only on some upper BEOL layers
- Remove seal ring on layers with wires
- FAB test structures must be moved to make way for wires
- Short wires enable high bandwidth at low cost and low power

Co-designed with
fab and chip architecture



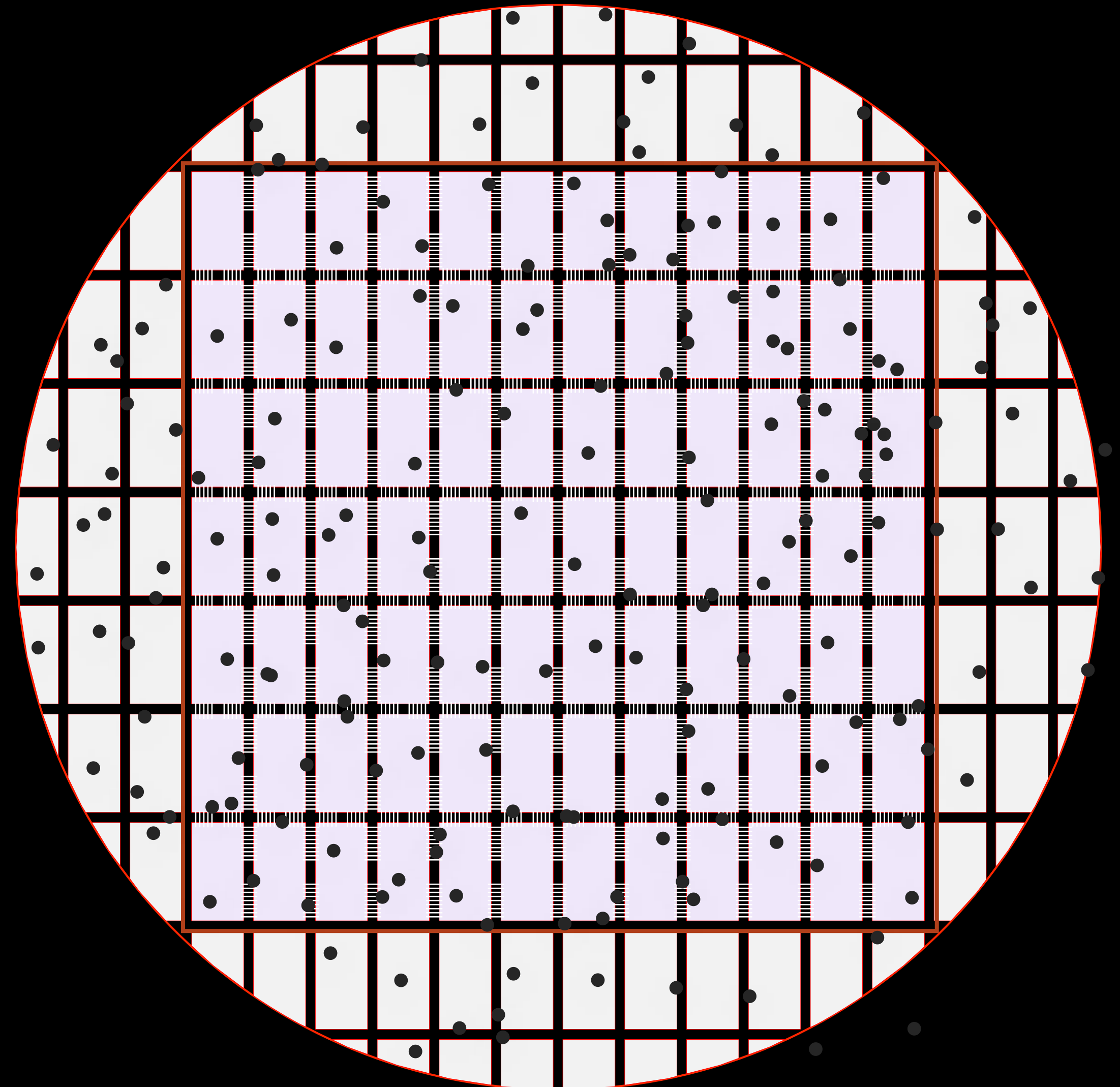
DEFECTS ARE EVERYWHERE

- It is impossible to yield a wafer with no defects

DEFECTS ARE UNAVOIDABLE

- Systematic defects are large and rare
- Most defects are stochastic, small and sparse
- Most of the area has no defects
- Exploits statistical nature of defects

Select architecture that is
resilient to defects



CHOOSE A SMALL FAILURE DOMAIN

Example: < 500µm x 500µm AI Core

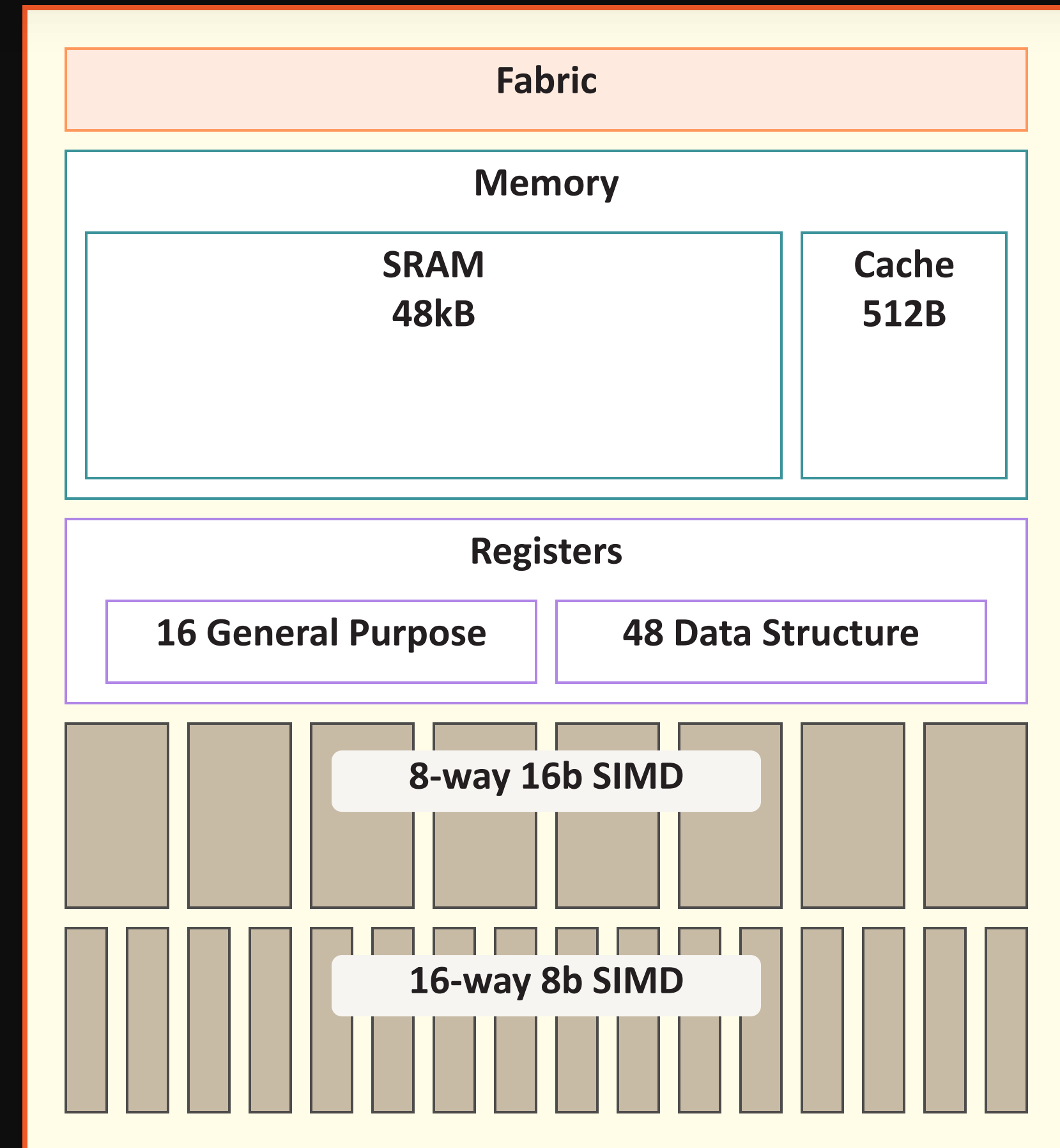
Improved performance for AI compute

- New higher performance tensor operations
 - New 8-way SIMD for 16b data (FP/BF16)
 - New 16-way SIMD for 8b data (Fixed/INT8)
- New faster non-linear functions
- 2x higher compute performance core

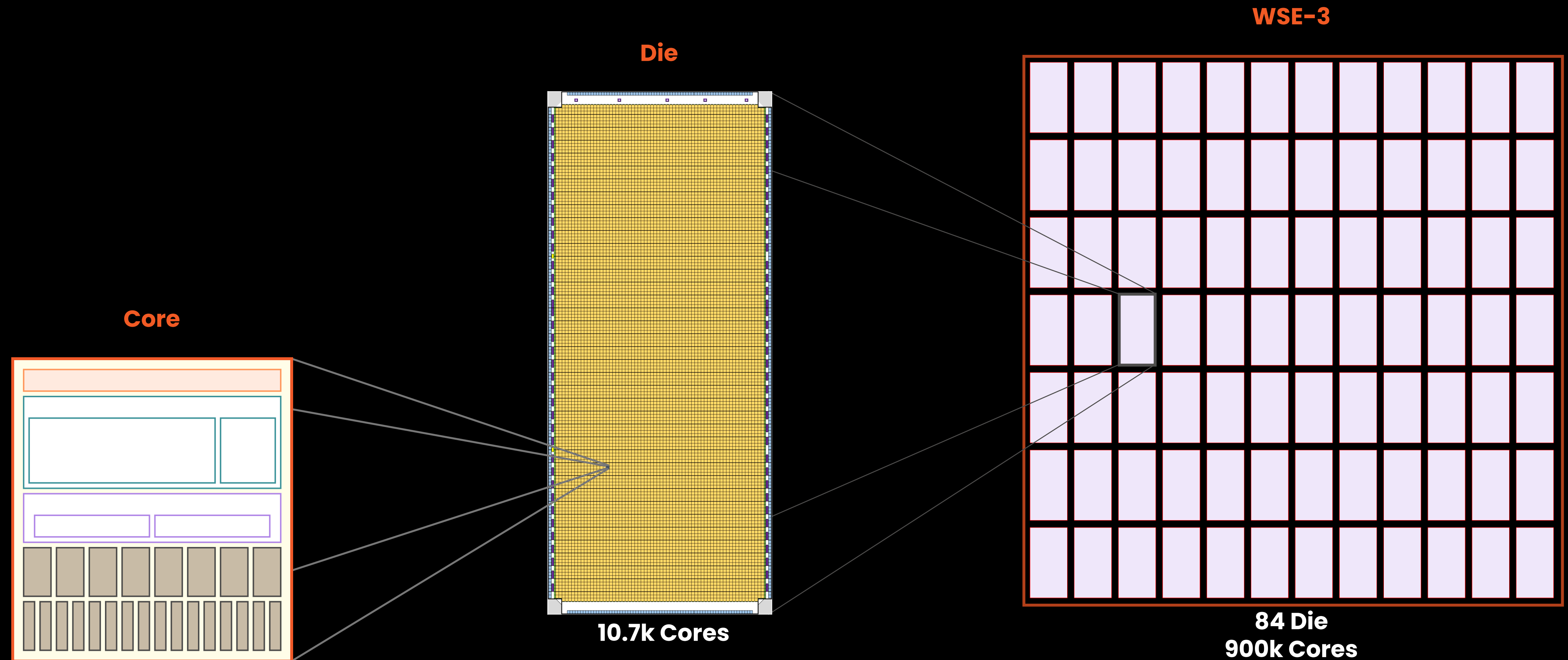
High bandwidth memory and cache

- 48kB memory per core
- New 512B local cache per core
- Full bandwidth for full SIMD performance

WSE-3 CORE

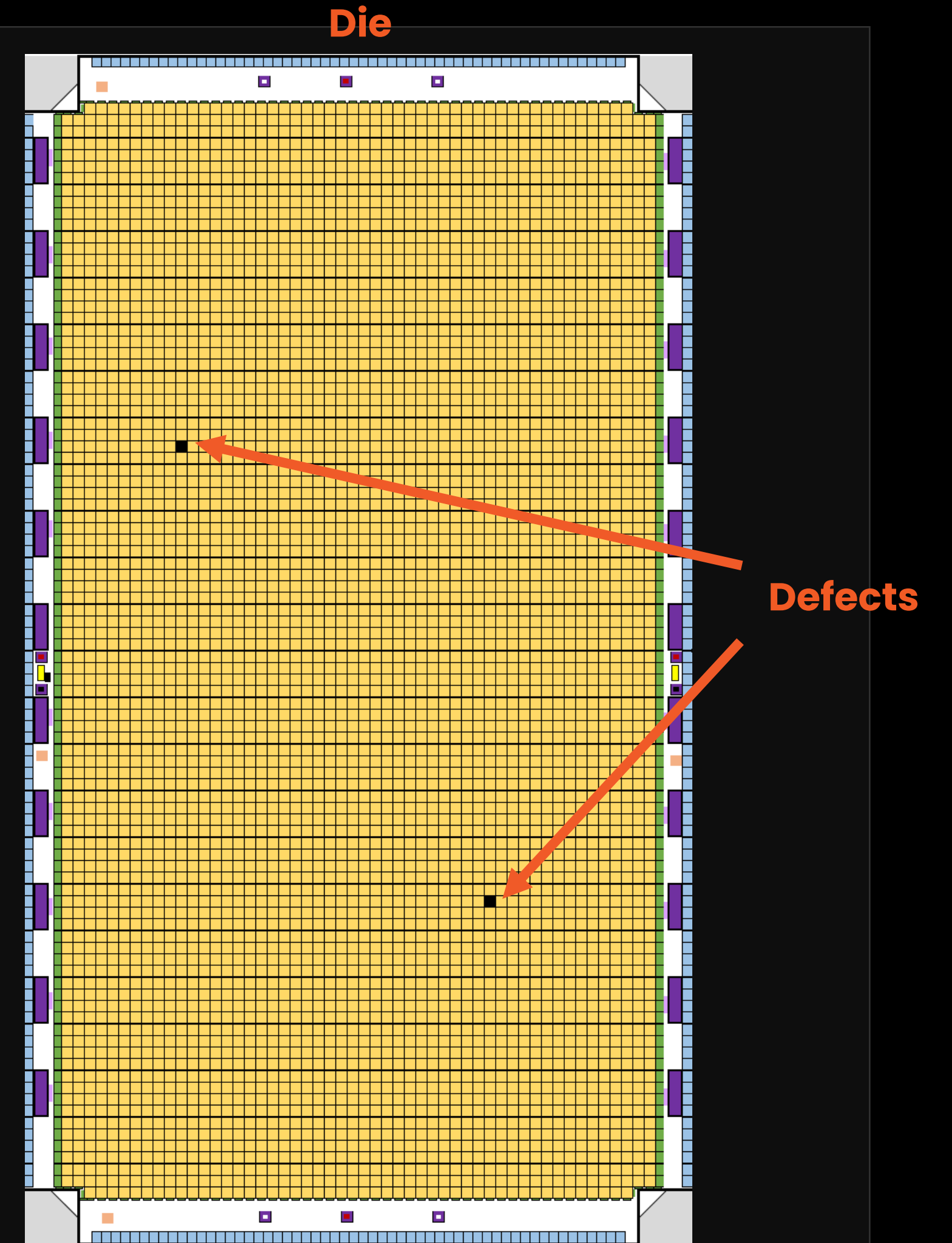


FROM SMALL CORE TO MASSIVE WAFER



SOLVING YIELD CHALLENGES

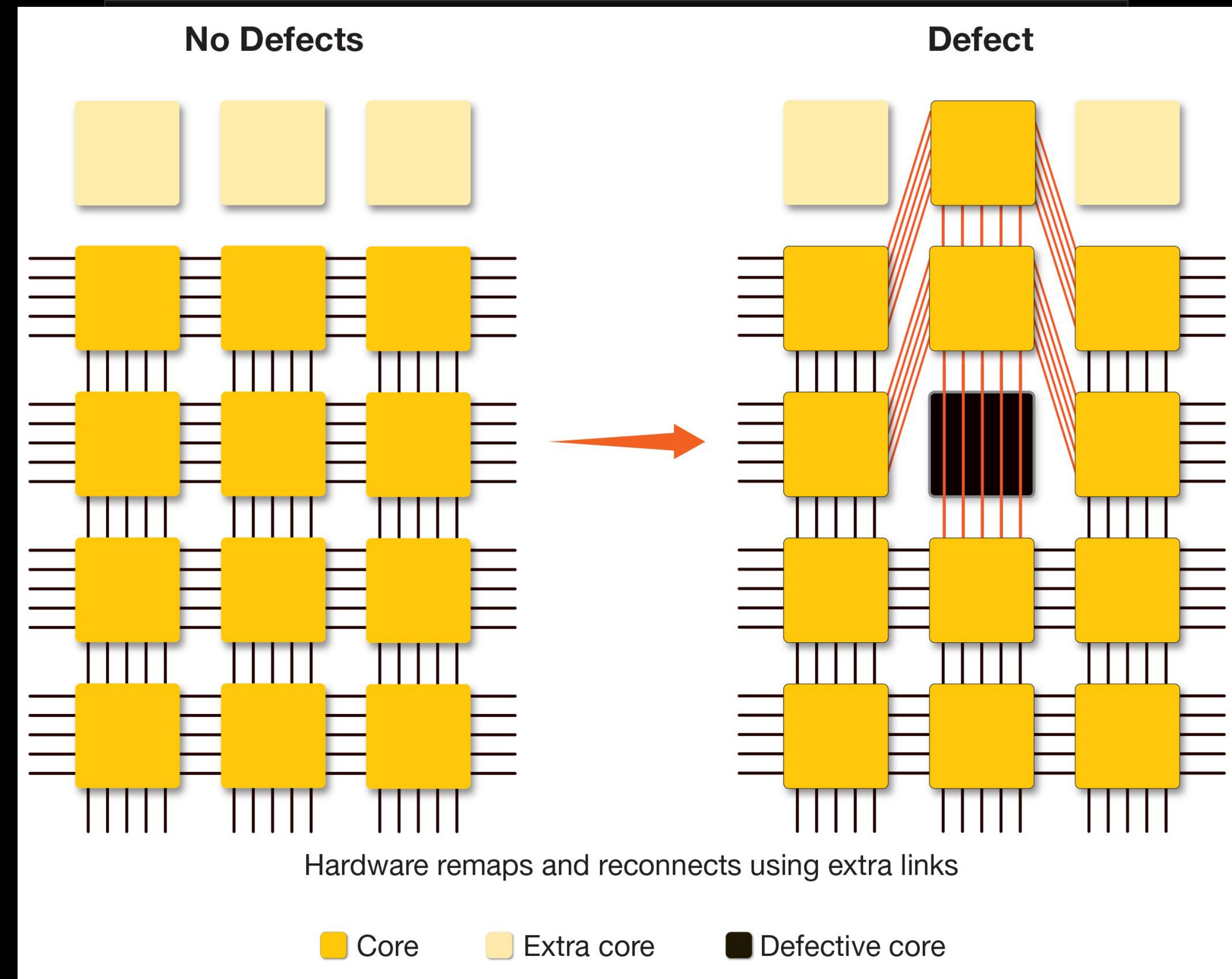
- Impossible to yield full wafer with zero defects
- Silicon and process defects are inevitable even in mature process



REDUNDANCY

- Uniform small core architecture enables redundancy to address yield at very low cost
- Design includes redundant cores and redundant fabric links
- Redundant cores replace defective cores
- Extra links reconnect fabric to restore logical 2D mesh

Co-designed with
chip architecture



HIDING DEFECTS TO UPPER SOFTWARE LAYERS

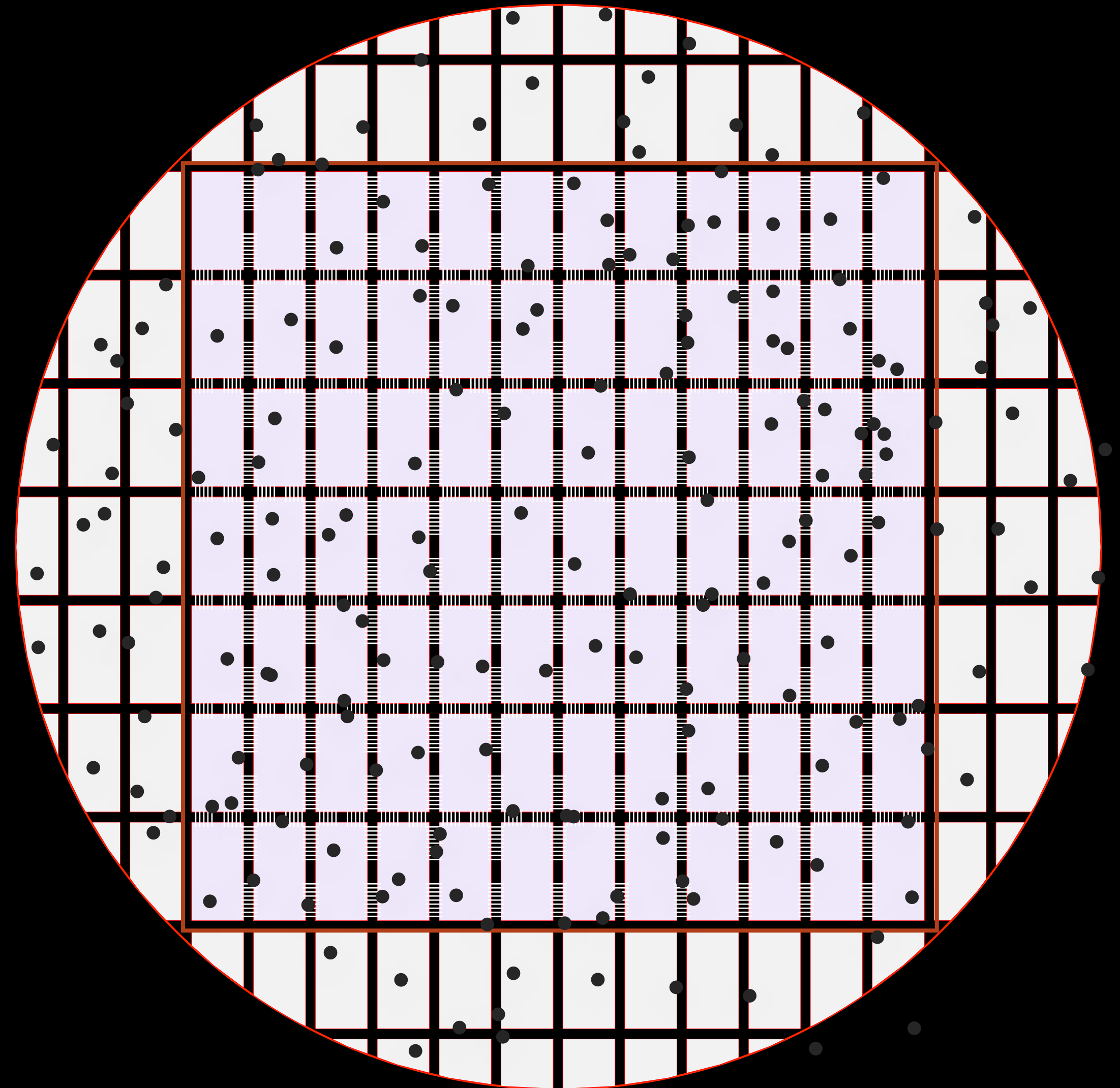
- Hierarchical software abstraction

LOW LEVEL SOFTWARE

- Detects defects
- Maps out defects
- Exposes pristine wafer to upper layers
- Tunes voltage to compensate for silicon

HIGH LEVEL SOFTWARE

- Unaware of defects
- Same wafer properties across different wafers



IMPROVING YIELD WITH FIELD REPAIRS

- Software-based repairs done in the field

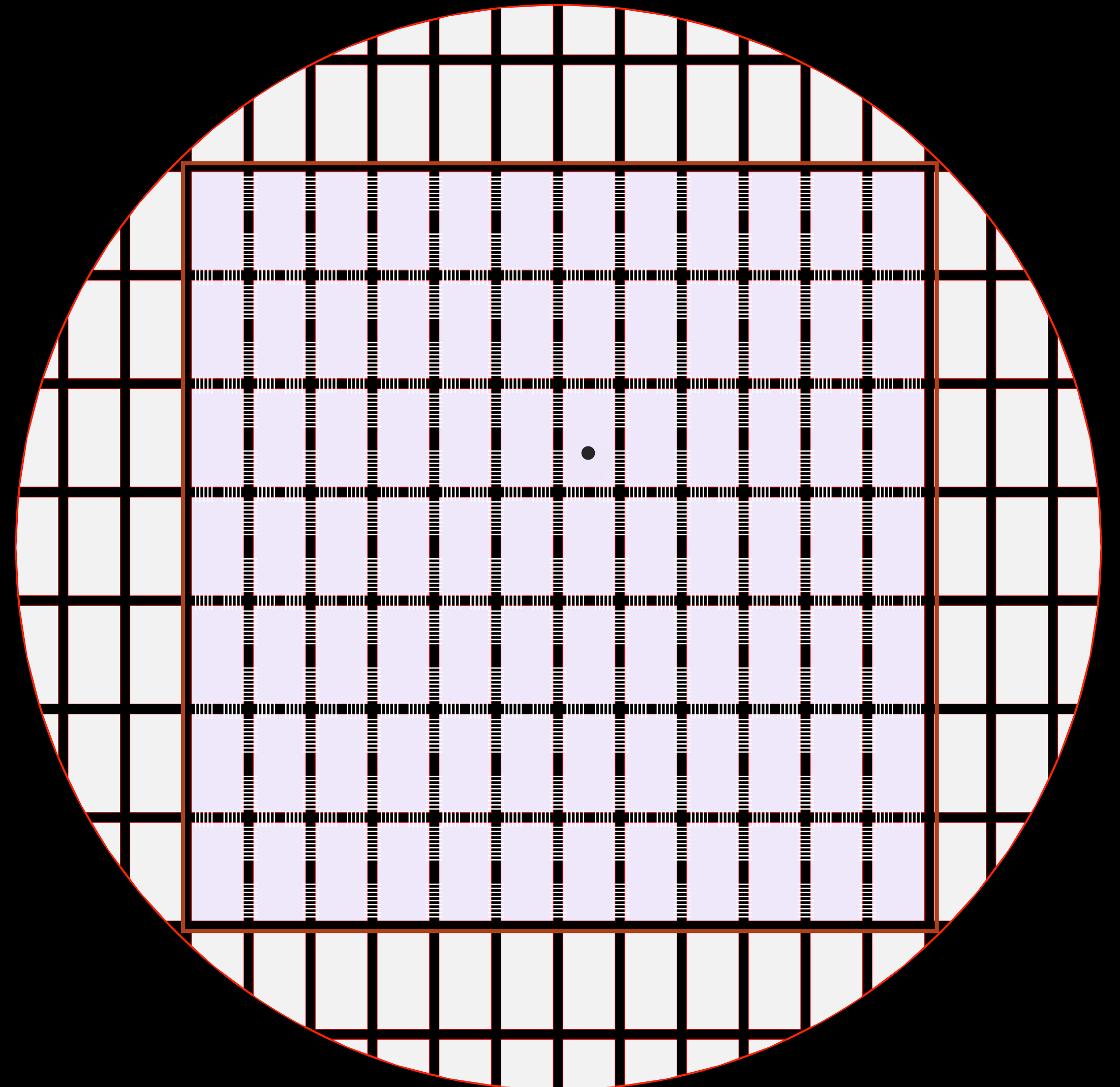
DEFECTS APPEAR OVER THE LIFETIME

- Silicon devices inherently fail over time

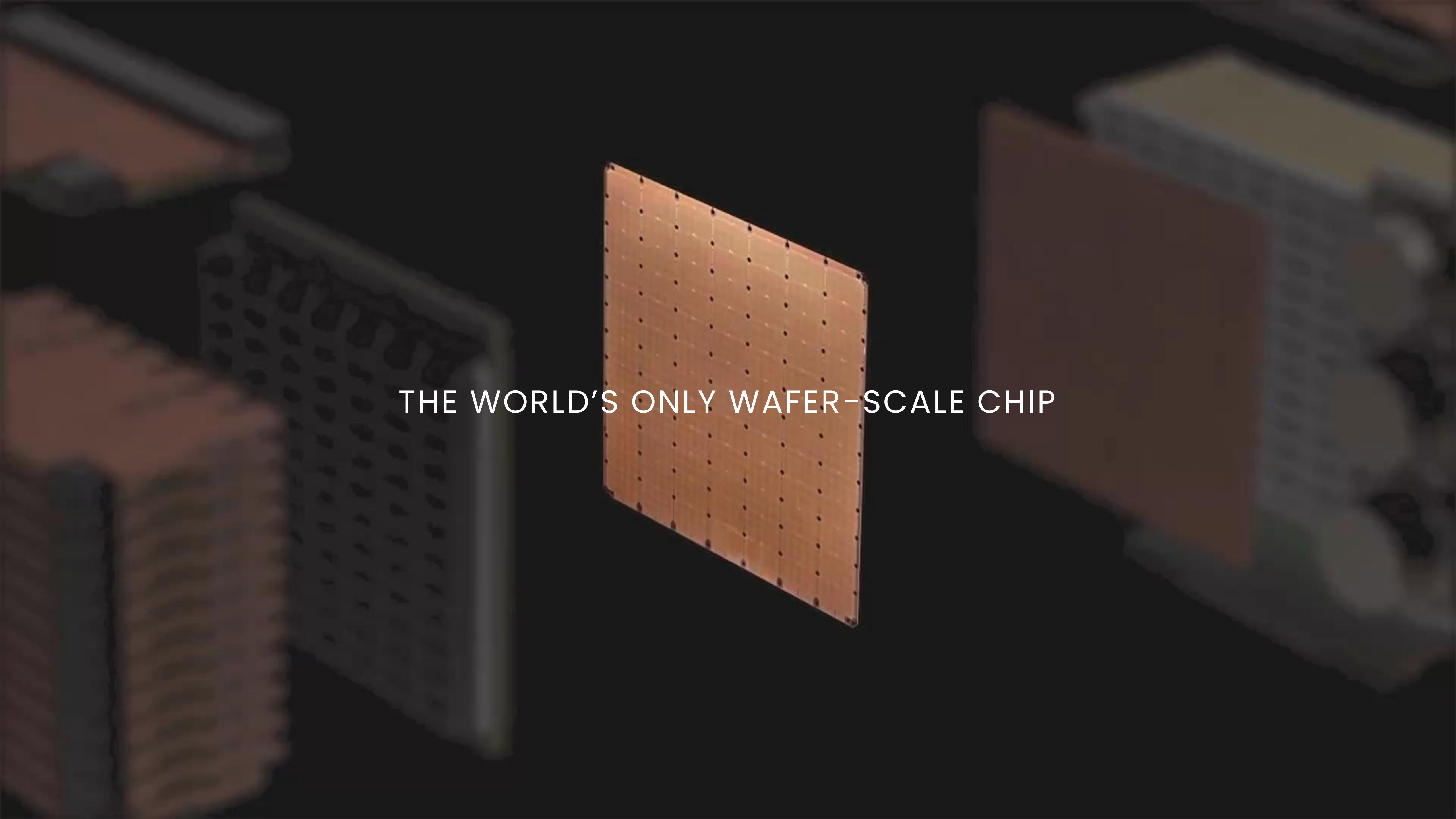
LOW LEVEL SOFTWARE

- Detects defects
- Maps out defects
- Exposes pristine wafer to upper layers

Extends lifetime without costly
component swaps



**WAFER SCALE IS NOT A LARGER CHIP
— IT IS A DIFFERENT SYSTEM-DESIGN DISCIPLINE**

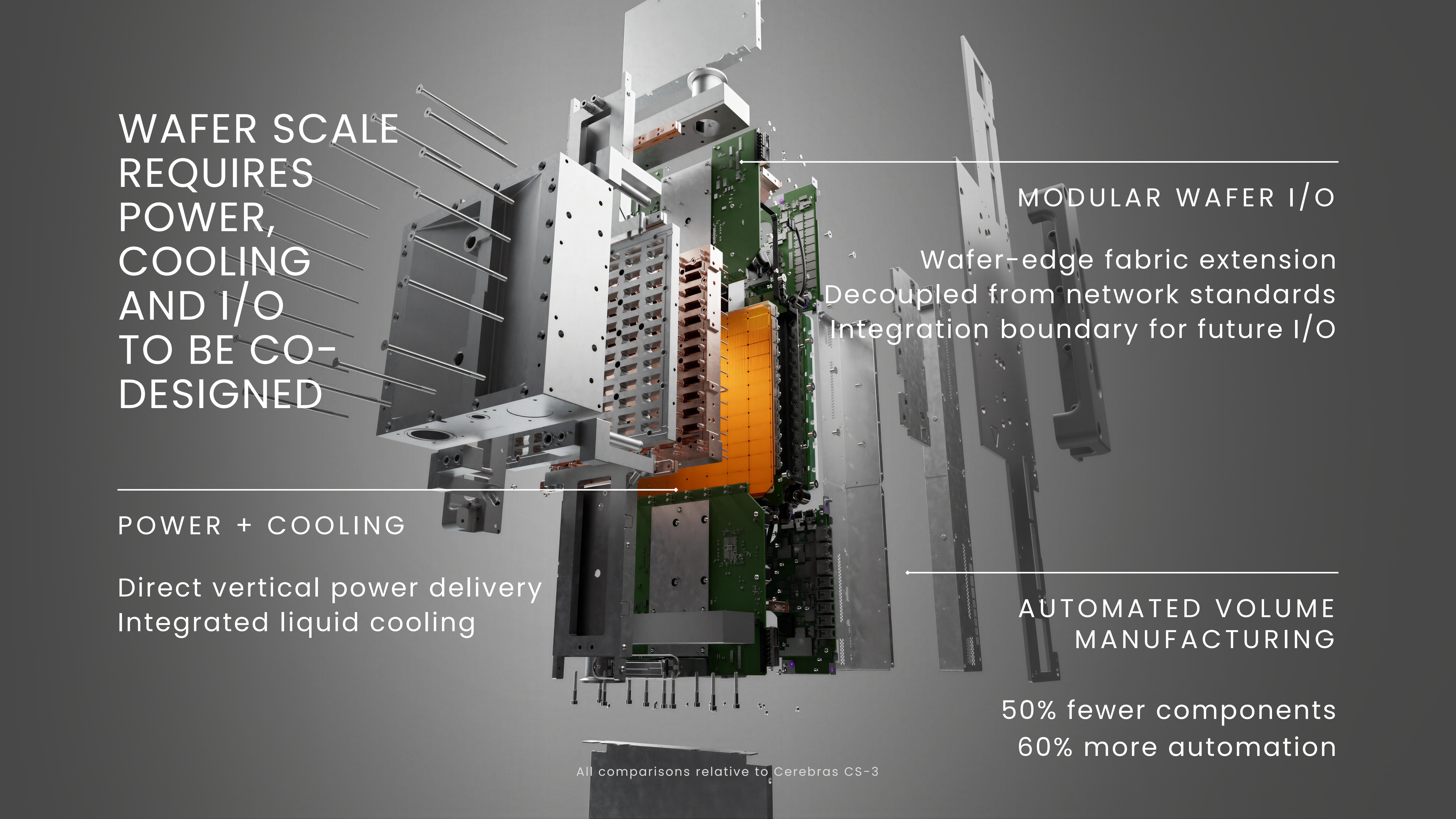


THE WORLD'S ONLY WAFER-SCALE CHIP



INTRODUCING

THE CEREBRAS CS-4



WAFAER SCALE
REQUIRES
POWER,
COOLING
AND I/O
TO BE CO-
DESIGNED

POWER + COOLING

Direct vertical power delivery
Integrated liquid cooling

MODULAR WAFER I/O

Wafer-edge fabric extension
Decoupled from network standards
Integration boundary for future I/O

AUTOMATED VOLUME MANUFACTURING

50% fewer components
60% more automation

All comparisons relative to Cerebras CS-3

MODULAR WAFER I/O DECOUPLES COMPUTE FROM CONNECTIVITY

PLUGGABLE OPTICS

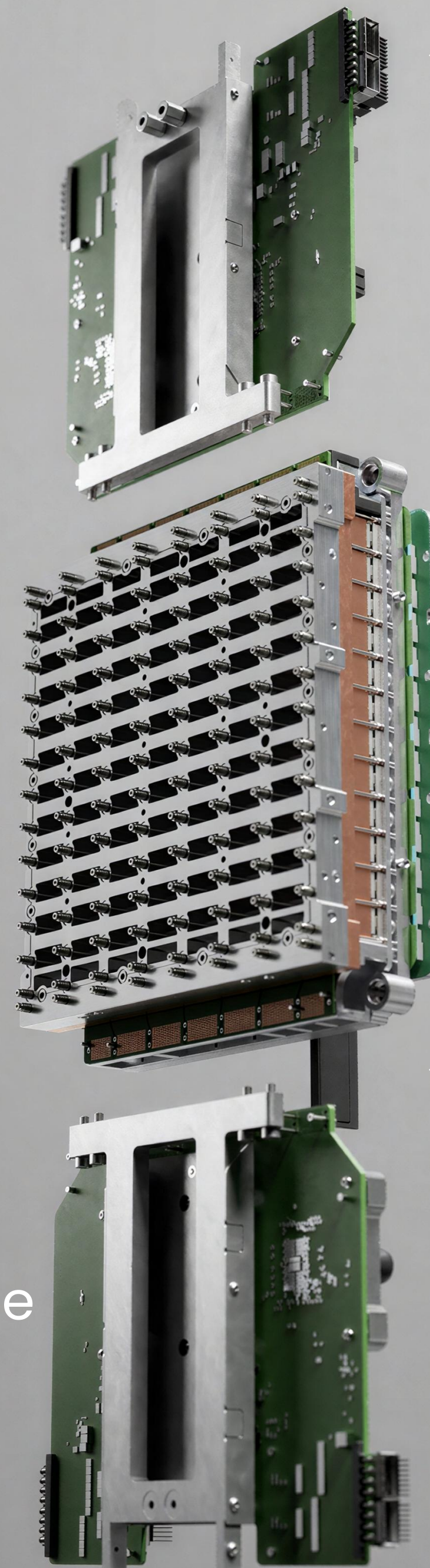
Standards-based RoCE networking
Multiple vendors
Serviceable

ARCHITECTURAL SEPARATION

Extends fabric from the wafer edge
Modular and programmable I/O
evolve independently of compute

TODAY: HIGH-SPEED ELECTRICAL I/O

Direct wafer-link interfaces



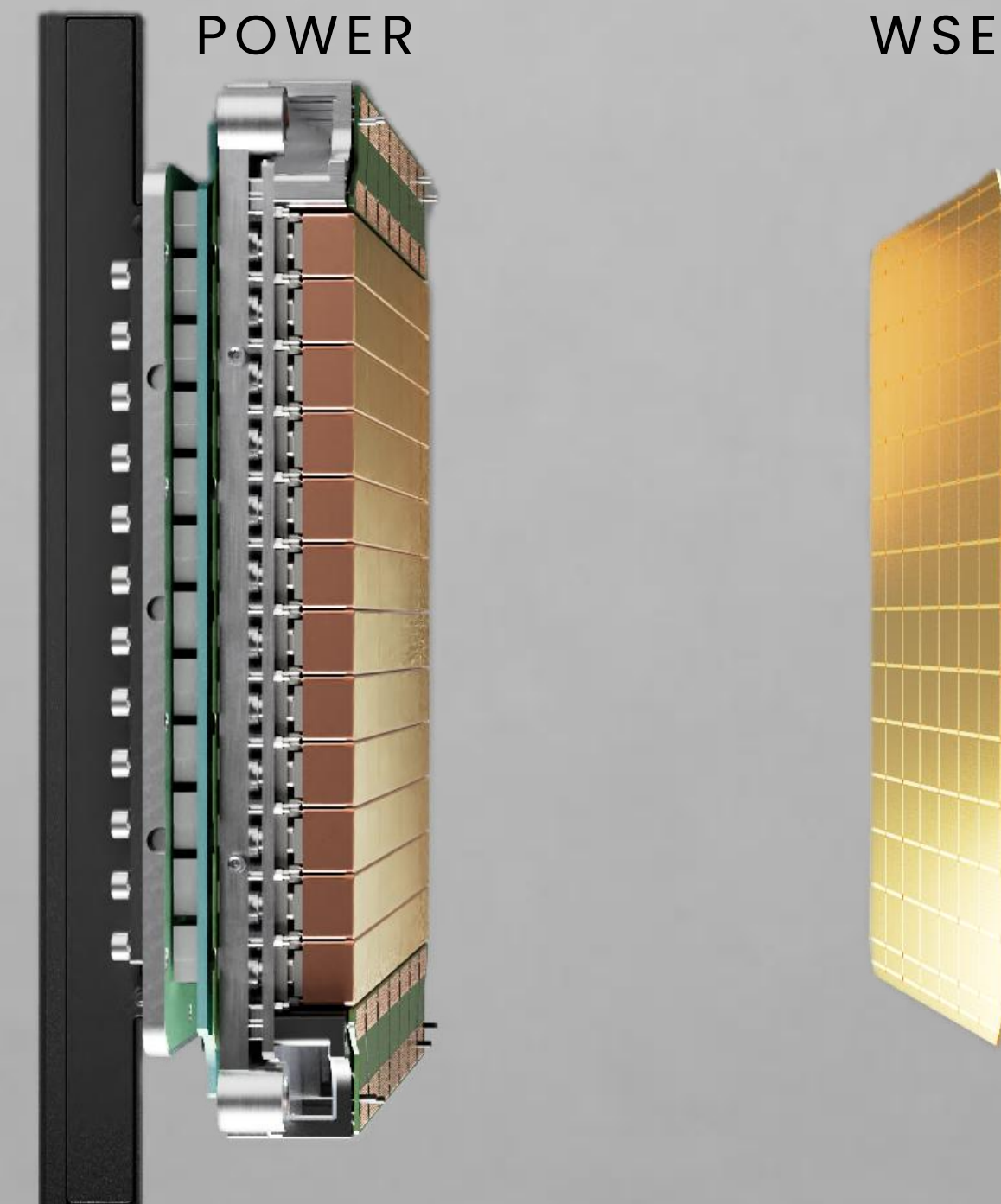
THE FUTURE OF HETEROGENOUS INTEGRATION

WAFER-SCALE BUILT OUR 3D INTEGRATION ADVANTAGE

Announced at HotCHIPS' 26

INDUSTRY LEADING
PRODUCTION PROVEN

YIELD-RESILIENT ARCHITECTURE
VERTICAL POWER DELIVERY
INTEGRATED COOLING



CS-6: CEREBRAS WAFER-SCALE SRAM, WITH 3D STACKED DRAM

Announced at HotCHIPS' 26

3D WAFER-SCALE

WORLD'S FASTEST INFERENCE SPEEDS
ORDER OF MAGNITUDE SMALLER FOOTPRINT

WAFER-SCALE SRAM SCALES

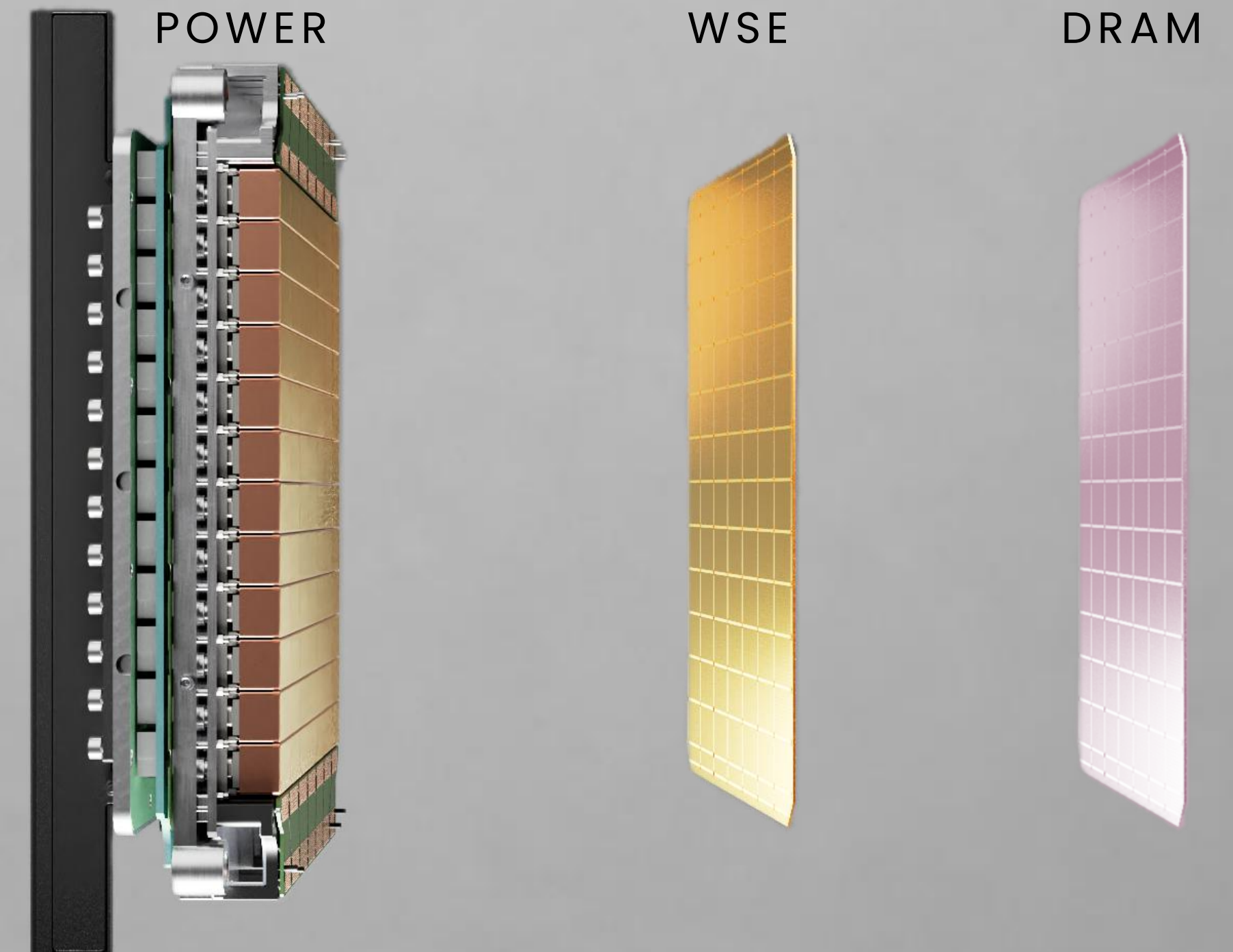
TO THE LARGEST MODELS

DRAM ENABLES SMALLER FOOTPRINT

LEVERAGING A DECADE OF

3D INTEGRATION INNOVATION

TO BRING ULTRAFAST INFERENCE TO EVERYONE



WAFER-SCALE SYSTEMS CREATED THE INTEGRATION FOUNDATION FOR HETEROGENEOUS HYBRID BONDING

Forward-looking example

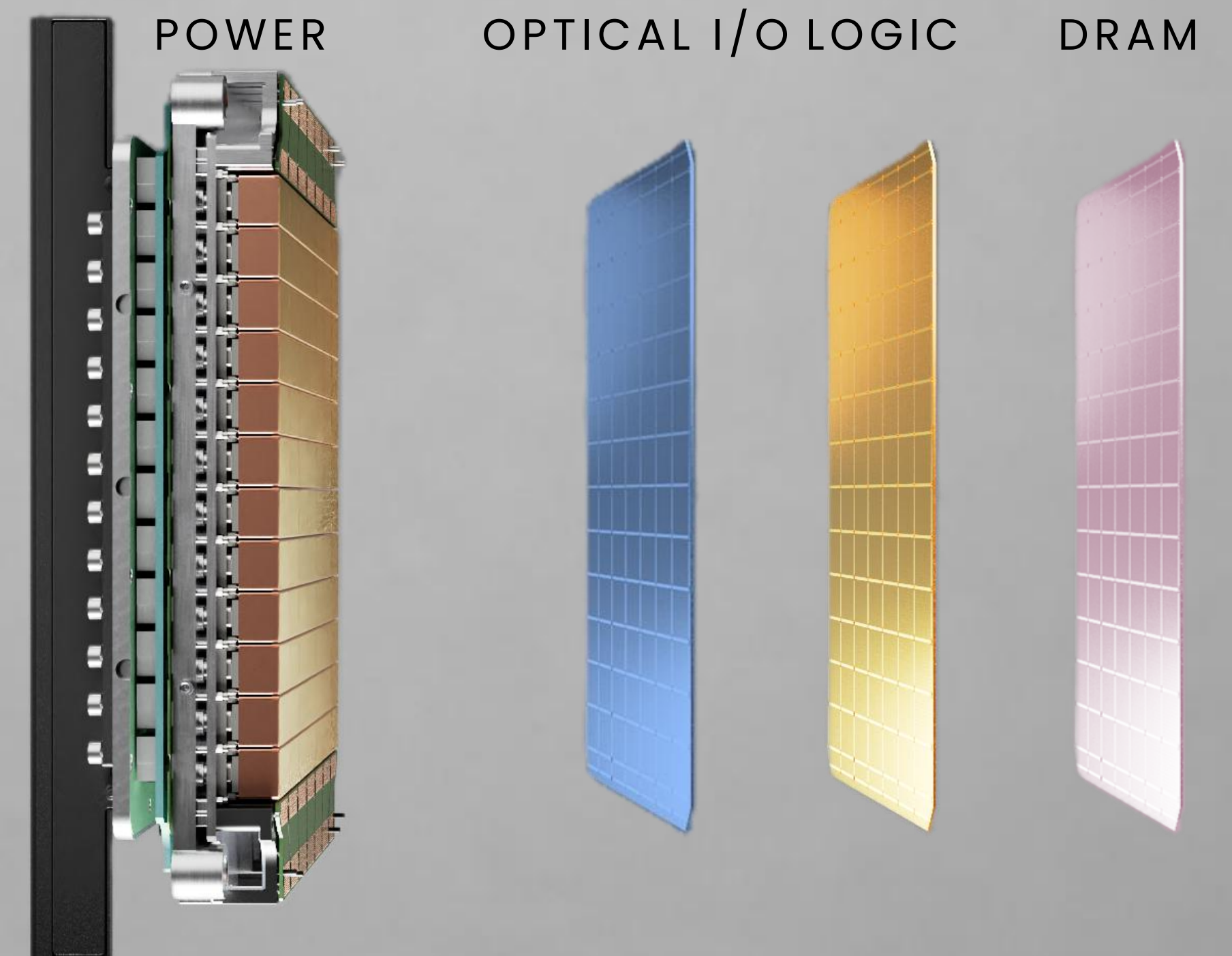
PRODUCTION-SCALE
YIELD-RESILIENT WAFER INTEGRATION

VERTICAL POWER DELIVERY

INTEGRATED COOLING

WAFER-LEVEL MECHANICAL AND THERMAL CONTROL

TEST AND MANUFACTURING INFRASTRUCTURE



Heterogeneous Hybrid Bonding extends integration capability from single compute wafer to dissimilar logic, memory and photonic technologies

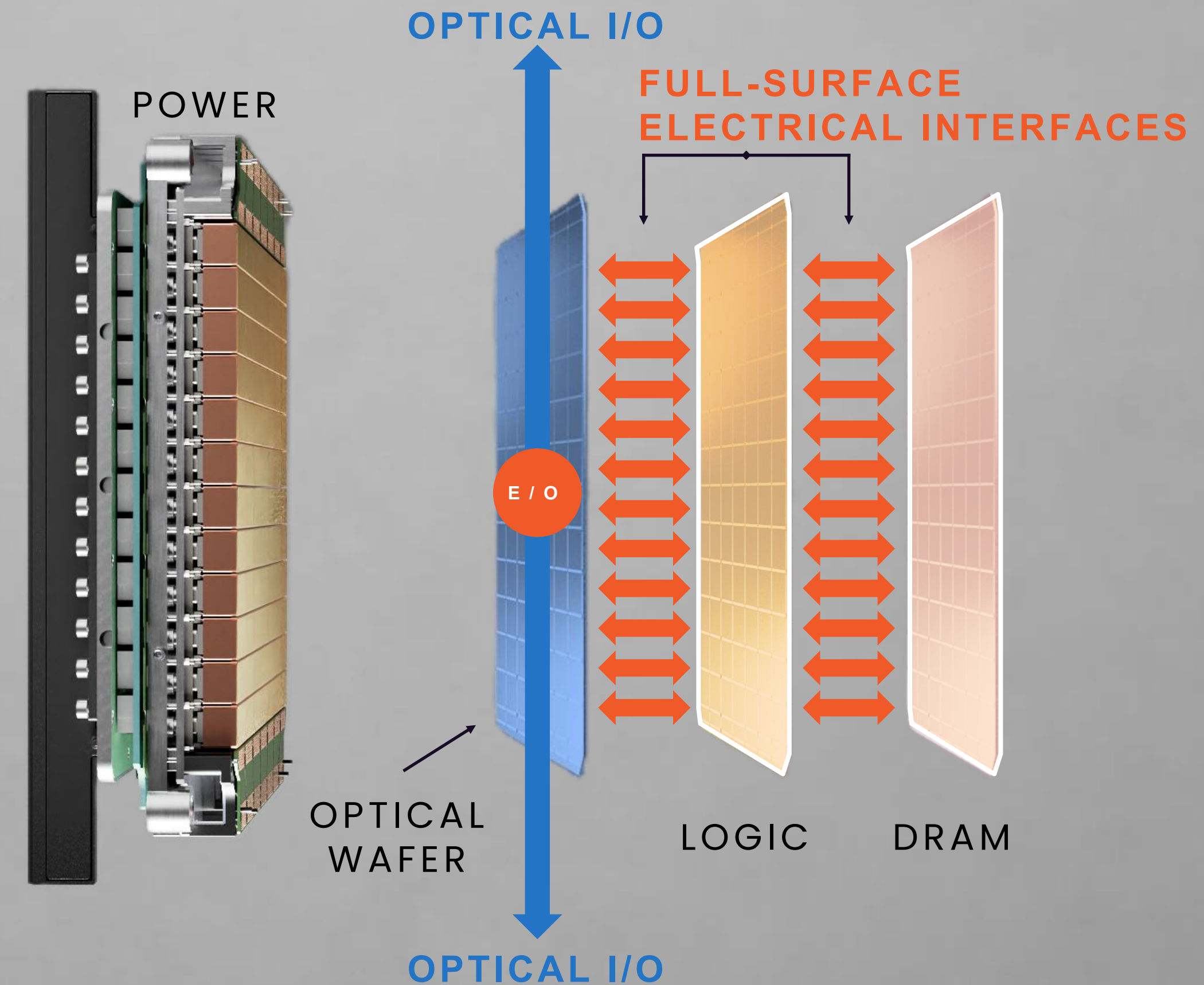
E/O CONVERSION TURNS WAFER-SURFACE BANDWIDTH INTO LOW-LOSS OPTICAL ESCAPE

Forward-looking example

DENSE ELECTRICAL INTERFACES
SPAN THE BONDED WSE SURFACE

E/O CONVERSION CAN OCCUR
NEAR EVERY COMPUTE REGION

LOW-LOSS WAVEGUIDES CARRY SIGNALS
FROM WAFER CENTER TO EDGES



Optical transport avoids the power losses and bandwidth-density limits of carrying electrical signals across the wafer

THE BONDED INTERFACE IS ONLY THE FIRST QUALIFICATION GATE

Pitch, overlay, planarity and bond defects must be solved — but they do not determine whether a heterogeneous wafer stack can be manufactured, yielded and powered at scale.

THE FAMILIAR BONDING QUESTIONS

CAN WE FORM A RELIABLE INTERFACE?

- Bond pitch and overlay tolerance
- Surface planarity and wafer bow
- Particle, void and open-defect control
- Bond strength and reliability

WHAT ACTUALLY DETERMINES SCALE

01 WAFER OWNERSHIP + TEST INSERTION

Where is each wafer tested, who certifies it as bond-ready, and what BEOL state preserves both test access and surface quality?

02 YIELD ARCHITECTURE ACROSS LAYERS

How does proven wafer-scale redundancy extend across heterogeneous layers so that defects do not compound through the stack?

03 STACK ORDER + VERTICAL FABRIC

Which wafer bonds first, and what uniform TSV fabric carries both power and signals through every layer?

THE STACK ARCHITECTURE AND MANUFACTURING FLOW MUST BE CO-DESIGNED

SCALING DEPENDS ON WHAT HAPPENS BETWEEN THE WAFERS — NOT JUST AT THE BOND

Each interface must arrive testable, each failure must stay local, and the completed stack must carry peak power and bandwidth.

01



02



03



KNOWN-GOOD-ENOUGH HANDOFF

Transfer custody only at a measured, bond-ready wafer interface.

BOUNDED FAILURE

Align repair regions so one defect costs capacity, not the stack.

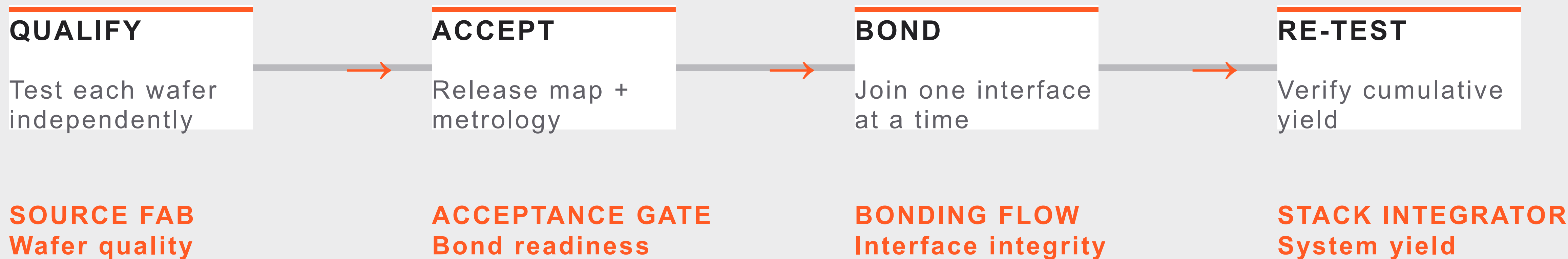
SHARED VERTICAL FABRIC

Set bond order around test, thermals, power, return paths and I/O.

THE THREE CONTRACTS MUST BE ARCHITECTED TOGETHER

OWNERSHIP TRANSFERS AT A TESTED, BOND-READY INTERFACE

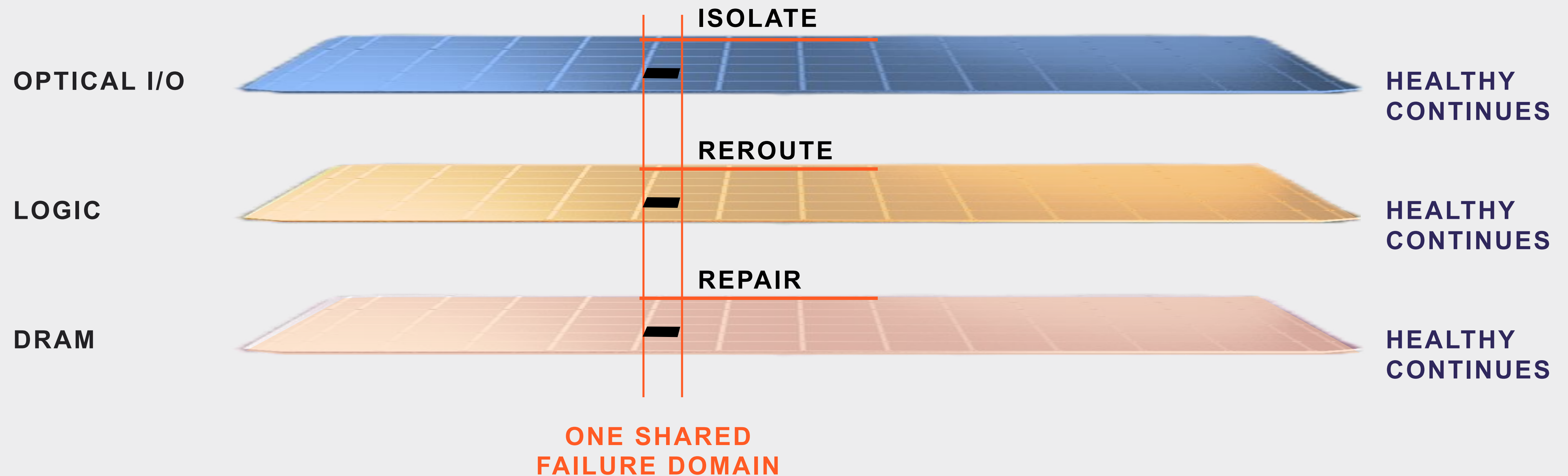
The acceptance package travels with the wafer: electrical map, surface metrology and explicit liability.



RESPONSIBILITY MOVES WITH MEASURED EVIDENCE — NOT WITH THE ORG CHART

ALIGNED REPAIR REGIONS KEEP DEFECTS LOCAL

One health map and consistent repair granularity let the stack route around a failed region while preserving healthy capacity.

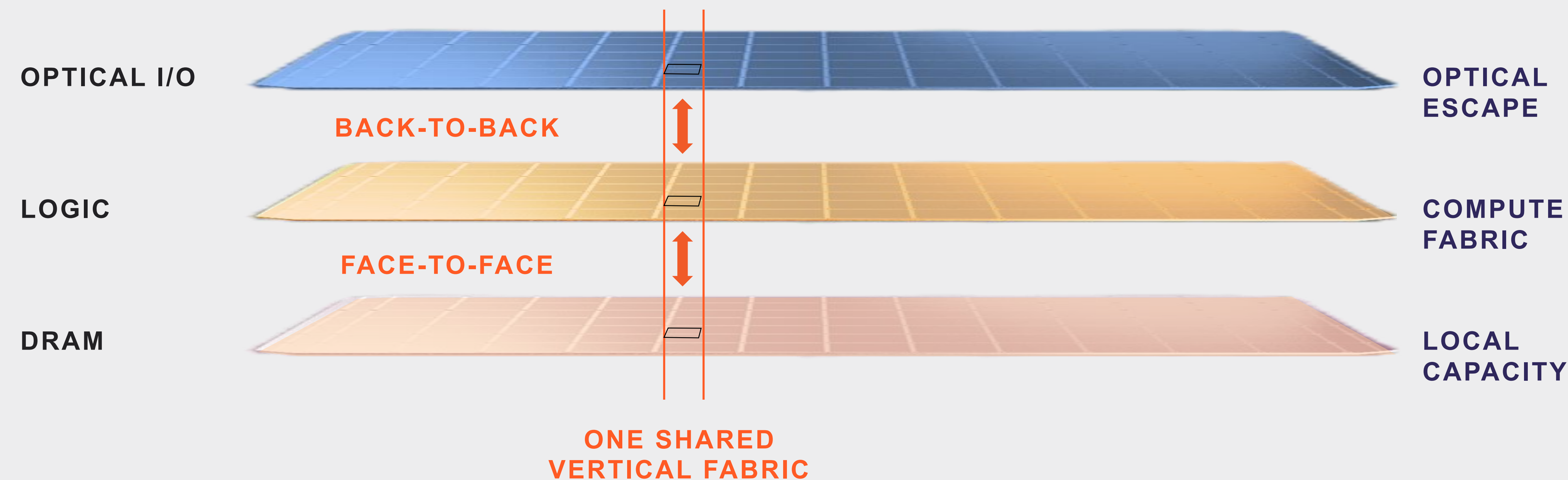


THE TARGET IS BOUNDED PERFORMANCE LOSS — NOT PERFECT WAFERS

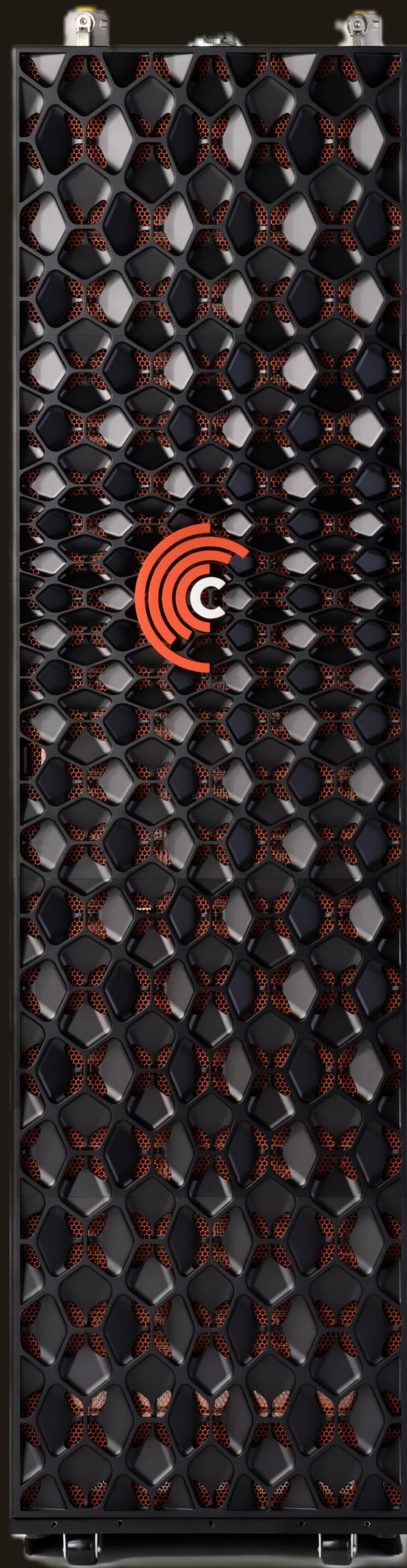
STACK ORDER

— SETS POWER, TEST AND BANDWIDTH LIMITS

Put optical I/O at the serviceable edge, keep logic close to DRAM, and size the shared vertical fabric for peak current and bandwidth.



BOND ORDER AND TSV TOPOLOGY ARE FIRST-ORDER ARCHITECTURAL PARAMETERS



HETEROGENEOUS HYBRID BONDING
SCALES WHEN THE STACK
IS ARCHITECTED
AS ONE SYSTEM

IT IS A DIFFERENT SYSTEM-DESIGN DISCIPLINE

Learn more

