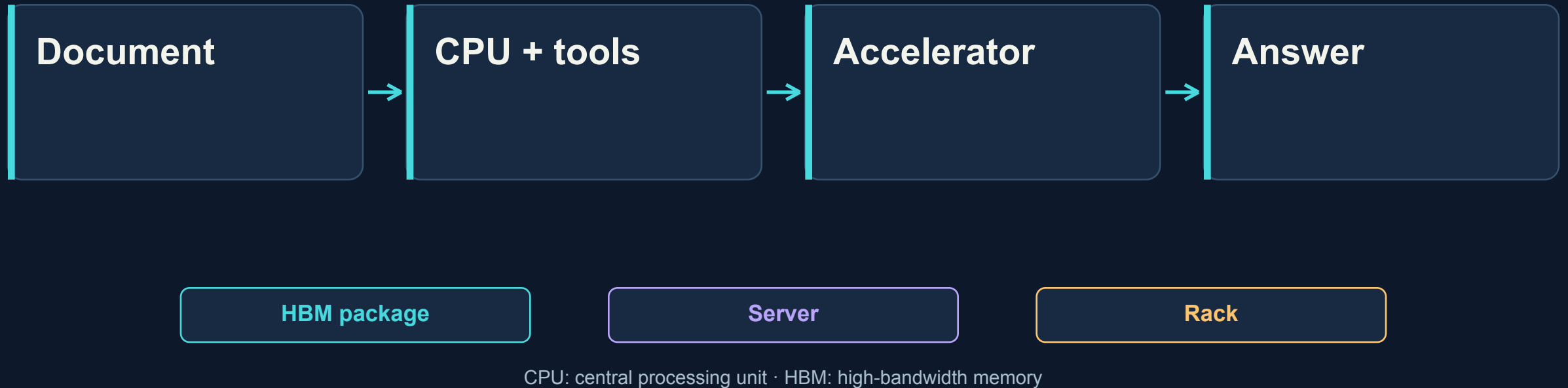


Start with the whole request

CHAPTER 01 / MEMORY CAPACITY AND THE REQUEST



Follow the data across every boundary

Original explanatory illustration · Sources: S04 pp3–4 · S01 pp5–9

The three memory questions

CHAPTER 01 / MEMORY CAPACITY AND THE REQUEST

Capacity

How much fits?

Warehouse space

Bandwidth

How fast in aggregate?

Loading dock

Latency

How long per access?

Delivery time

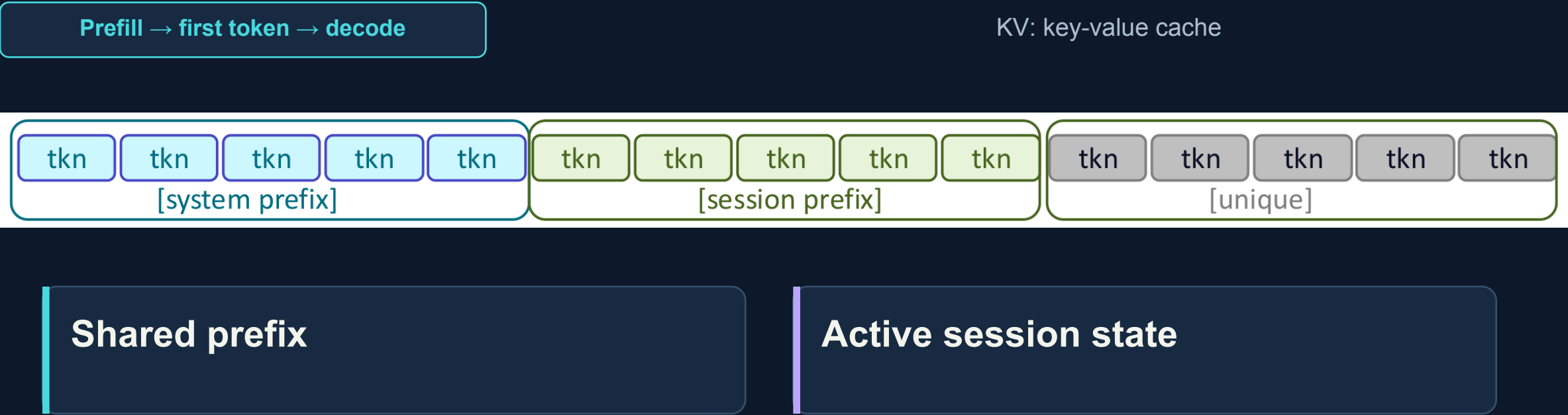
Match all three to the request pattern

Original explanatory illustration · Sources: S01 pp4–9



What the KV cache remembers

CHAPTER 01 / MEMORY CAPACITY AND THE REQUEST

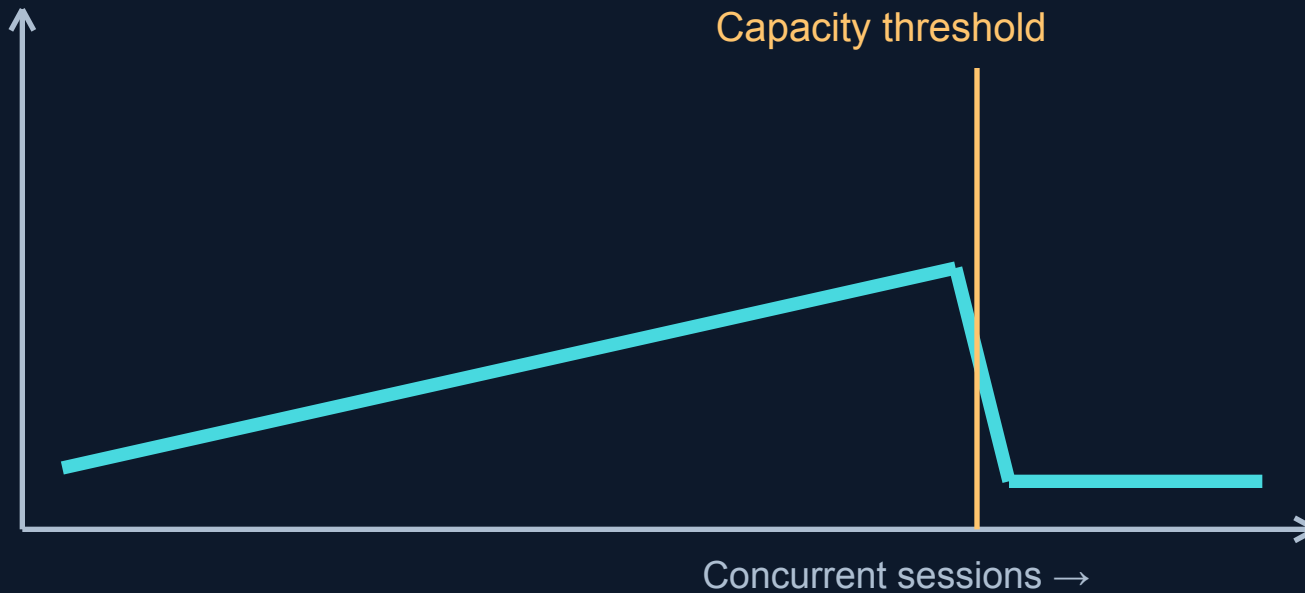


Different reuse patterns need different placement policies

The capacity cliff

CHAPTER 01 / MEMORY CAPACITY AND THE REQUEST

Useful serving capacity



$$2 \text{ GB} \times 50 = 100 \text{ GB}$$

60 sessions exceed capacity

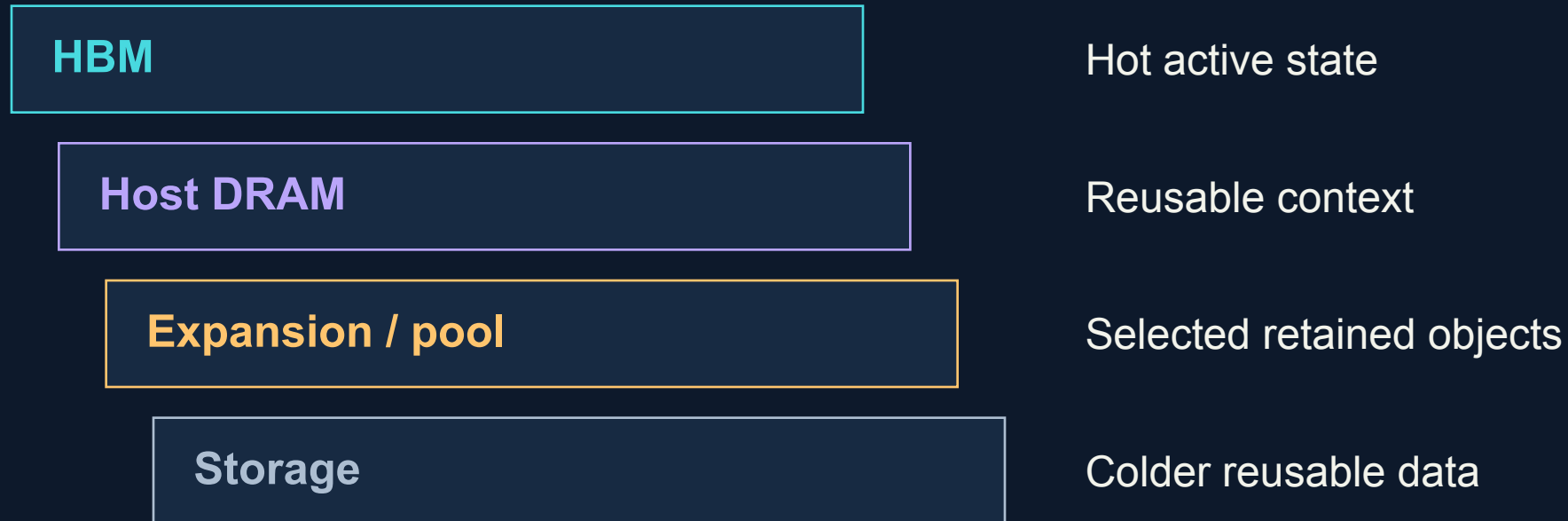
Illustrative arithmetic; excludes weights and runtime overhead

Original explanatory illustration · Sources: S01 pp10–13 · calculation original



Offload is a placement decision

CHAPTER 02 / PLACEMENT AND THE DATA PATH



Retain, retrieve, or recompute: placement is workload-dependent

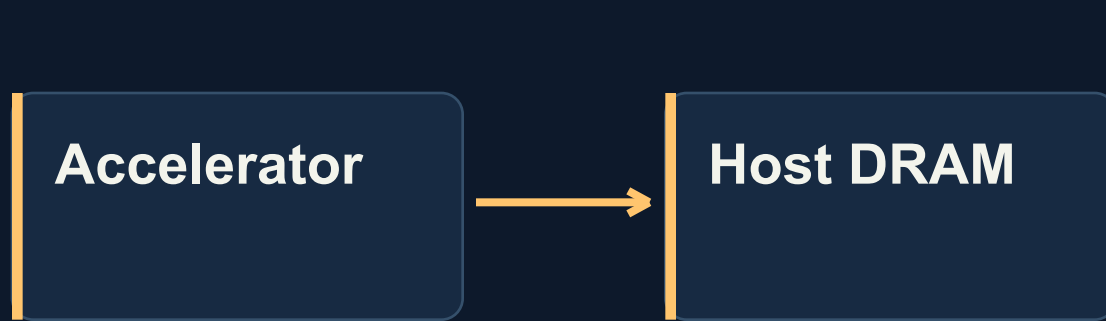
Original explanatory illustration · Sources: S01 pp8–16



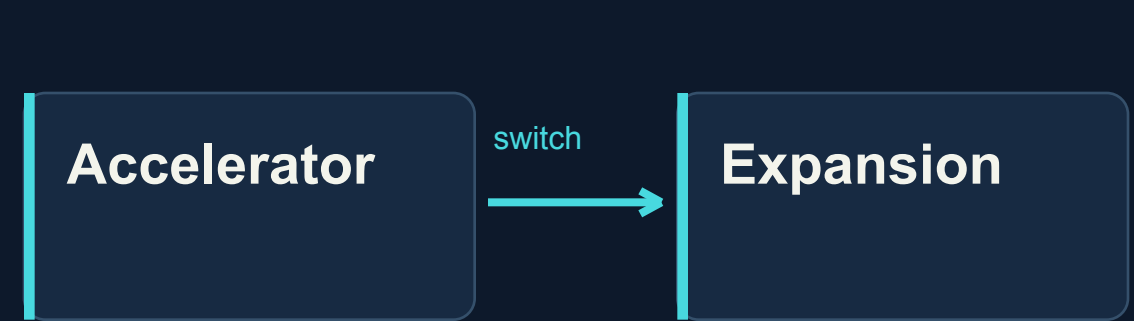
The hidden CPU detour

CHAPTER 02 / PLACEMENT AND THE DATA PATH

Host-memory route



Expansion route



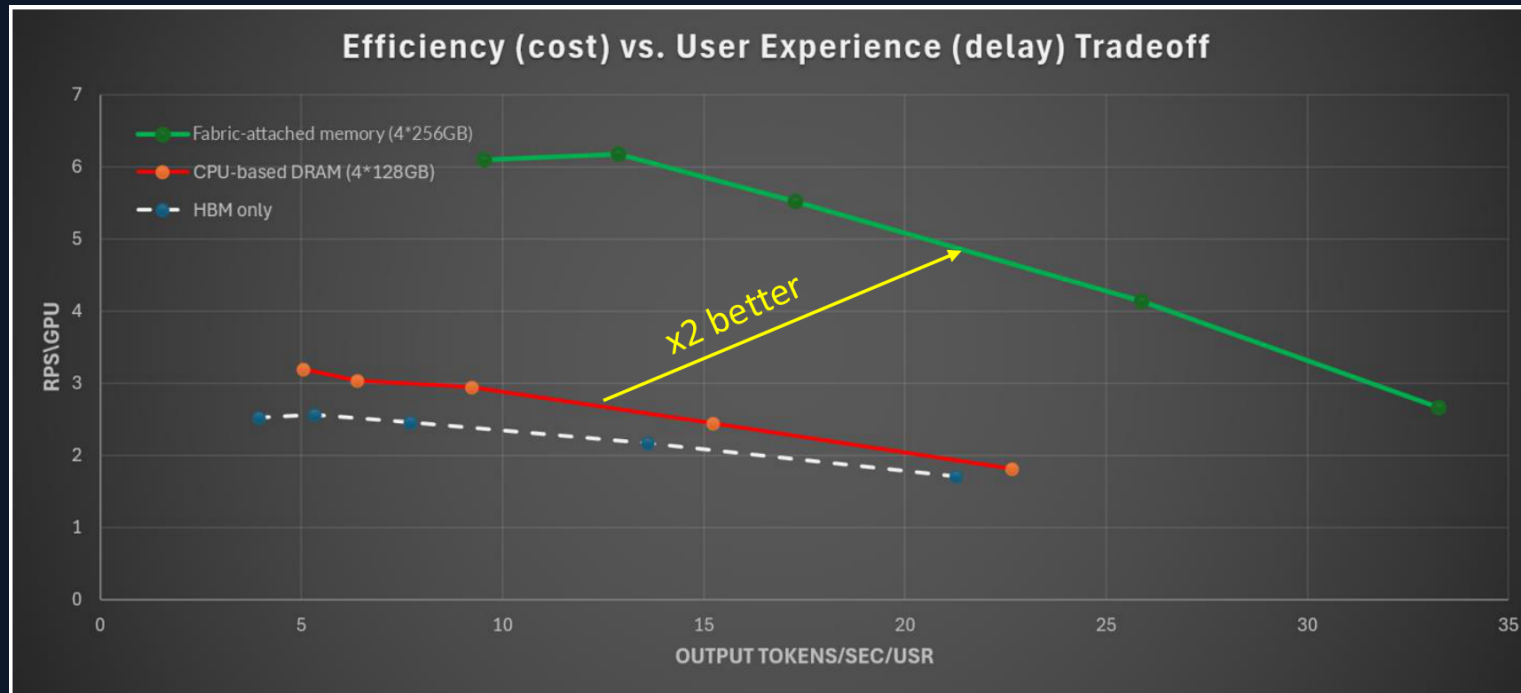
CPU control stays present

Capacity and the payload route are separate choices

Original explanatory illustration · Sources: S02 pp2–3

What the Astera experiment teaches

CHAPTER 02 / PLACEMENT AND THE DATA PATH



Parameters:

- Medium input prompts
- 4*128GB vs. 4*256GB

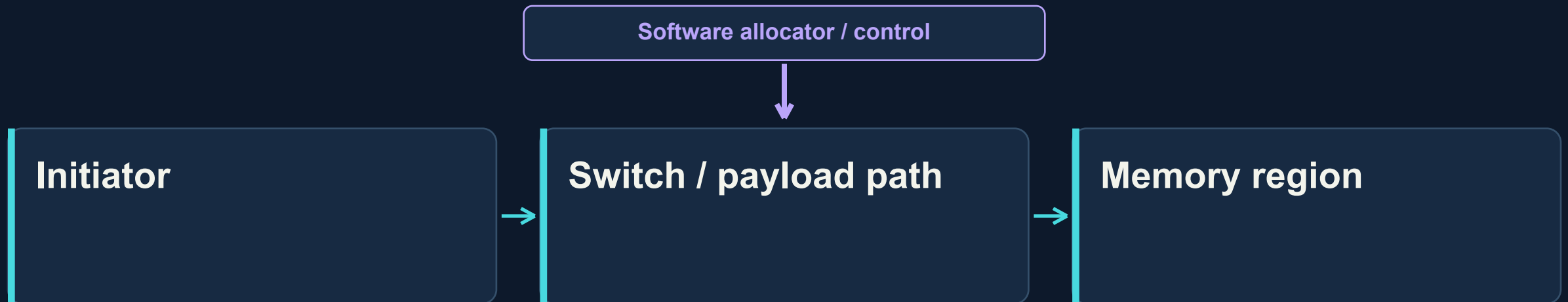
Qwen3-14B multi-turn
Benefit includes capacity difference

Short / equal: similar
Medium / larger: helps
Long / tiers: enables

The baseline has unequal capacity: 4×128 GB versus 4×256 GB

Draw the traffic before comparing products

CHAPTER 02 / PLACEMENT AND THE DATA PATH



Units: GB = capacity Gb/s = bit rate tokens/s = serving output

PCIe: Peripheral Component Interconnect Express · CXL: Compute Express Link

PCIe peer-to-peer ≠ GPU-native CXL memory semantics

Original explanatory illustration · Sources: S02 p5 · S06 pp14–15

Expansion pooling and sharing are different

CHAPTER 03 / SMART MEMORY AND AGENT PIPELINES

Expansion

More capacity
for one host

Pooling

Assign regions
from a shared pool

Sharing

Common object
+ access coordination

Access rules and ownership must be explicit

Original explanatory illustration · Sources: S05 · S05A

Move computation closer to the bytes

CHAPTER 03 / SMART MEMORY AND AGENT PIPELINES

Move the bytes to compute



Move the operation near memory

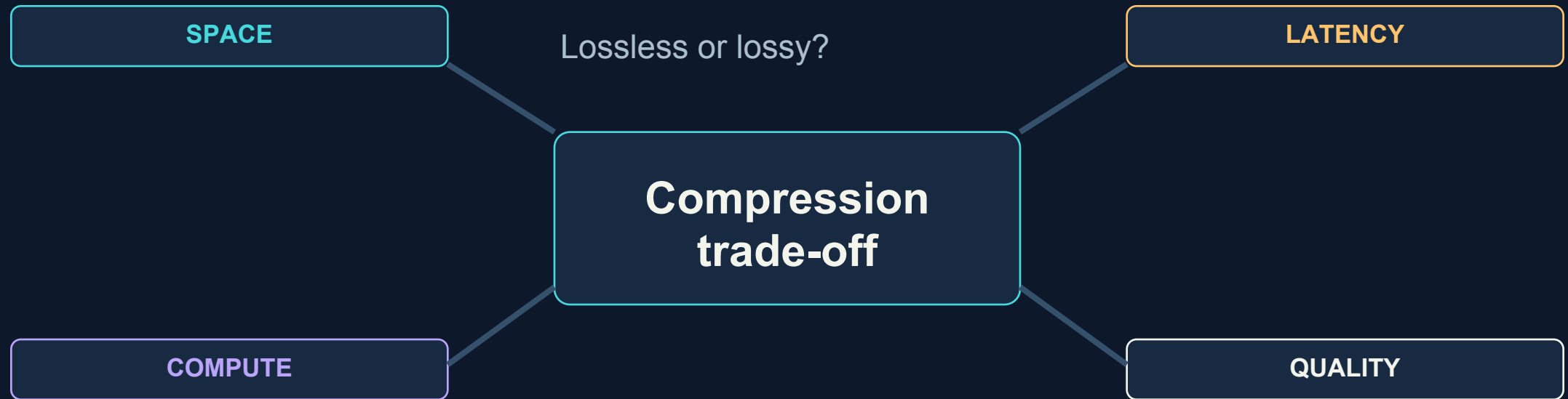


Include processing, transfer, and quality costs

Original explanatory illustration · Sources: S06 pp7–11 · S05B

Compression has a quality budget

CHAPTER 03 / SMART MEMORY AND AGENT PIPELINES



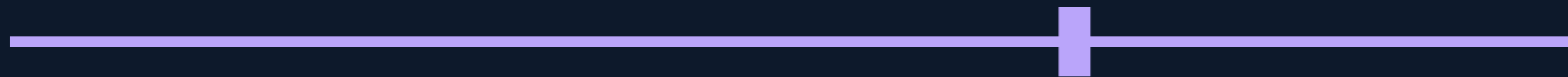
Optimize requests meeting quality AND latency targets

Original explanatory illustration · Sources: S06 pp11–12



The agent is a pipeline

CHAPTER 03 / SMART MEMORY AND AGENT PIPELINES



Stricter per-user generation target → fewer simultaneous users

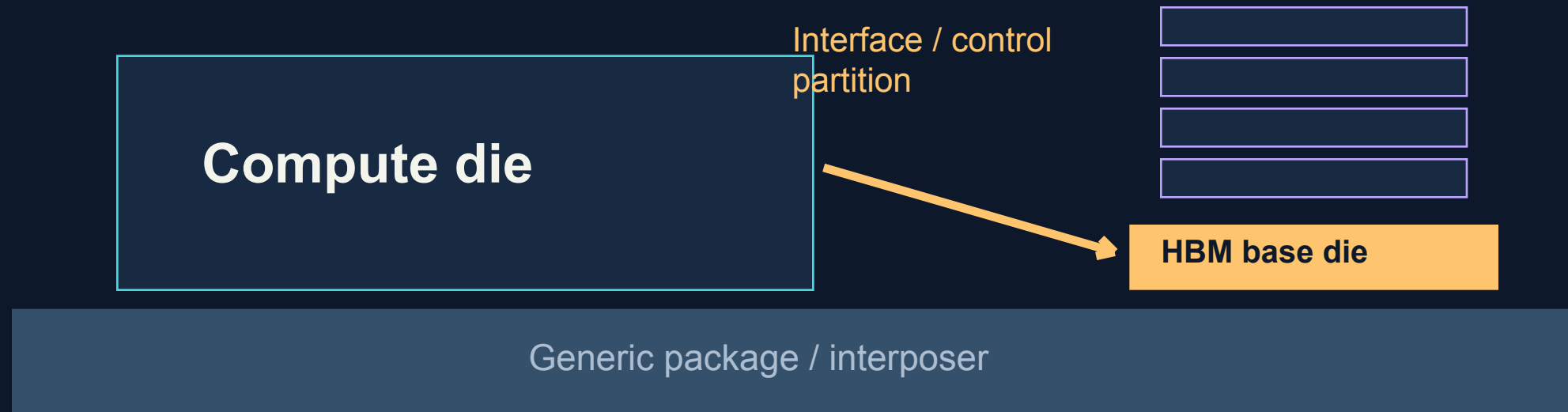
Measure the complete power boundary; benchmark host matters

Arm / FuriosaAI · Illustrated serving benchmark used an x86 host. Joint Arm-host validation was separate. [S04]



Why custom HBM changes the design boundary

CHAPTER 04 / PACKAGE AND THERMAL CO-DESIGN



Moving functions moves constraints

Original explanatory illustration · Sources: S03 pp3–4,12–14

Energy per bit becomes watts

CHAPTER 04 / PACKAGE AND THERMAL CO-DESIGN

$$1 \text{ pJ/bit} \times 1 \text{ Tb/s} = 1 \text{ W}$$

$$1 \text{ pJ/bit} \times 10 \text{ Tb/s} = 10 \text{ W}$$

$$10^{-12} \text{ J/bit} \times 10^{12} \text{ bit/s} = \text{J/s}$$

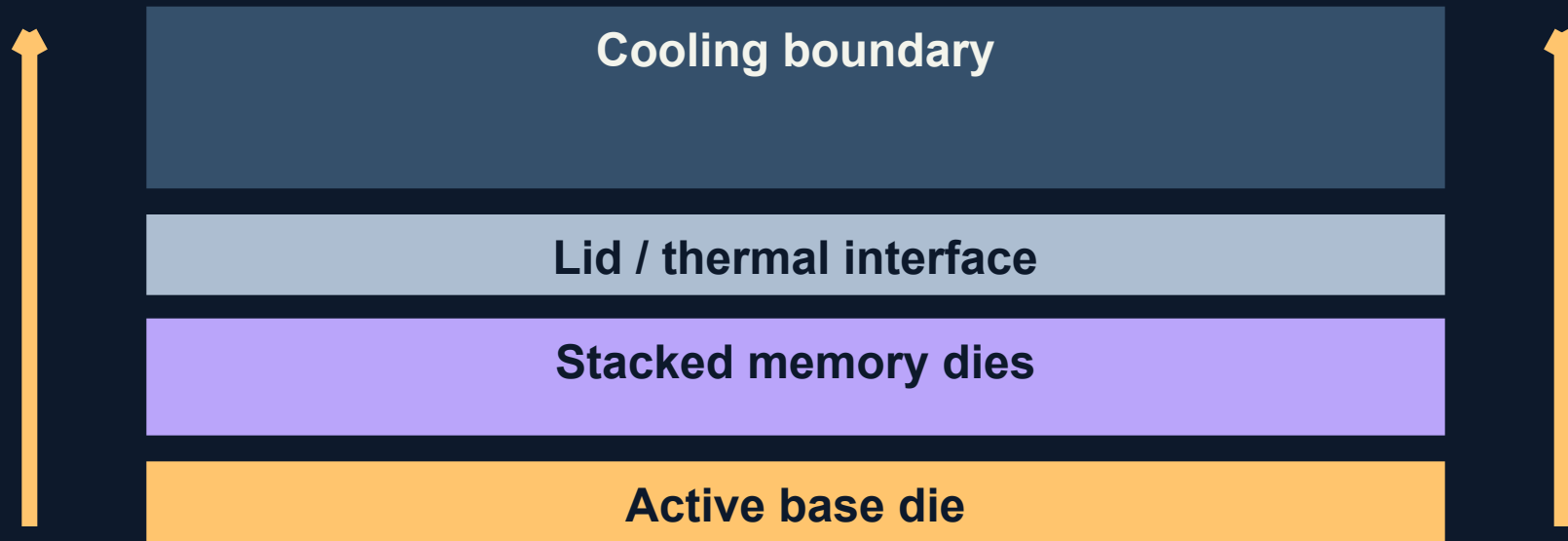
Illustrative arithmetic: define traffic pattern and power boundary

Original explanatory illustration · Sources: S03 pp10–11 · original dimensional calculation



Heat takes a physical path

CHAPTER 04 / PACKAGE AND THERMAL CO-DESIGN

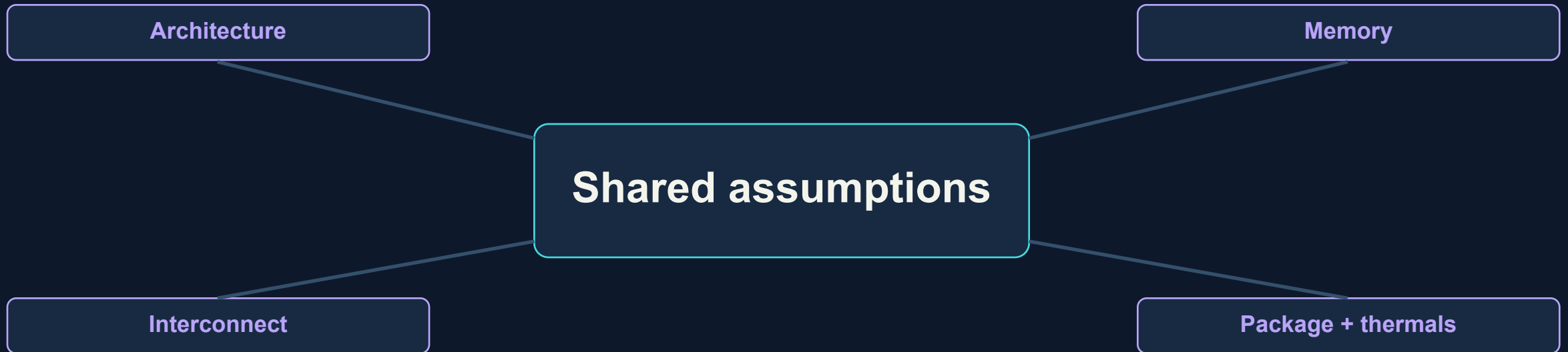


Watts + geometry + interfaces → temperature

Original explanatory illustration · Sources: S03 pp5–9 · generic original cross-section

The co-design review

CHAPTER 04 / PACKAGE AND THERMAL CO-DESIGN



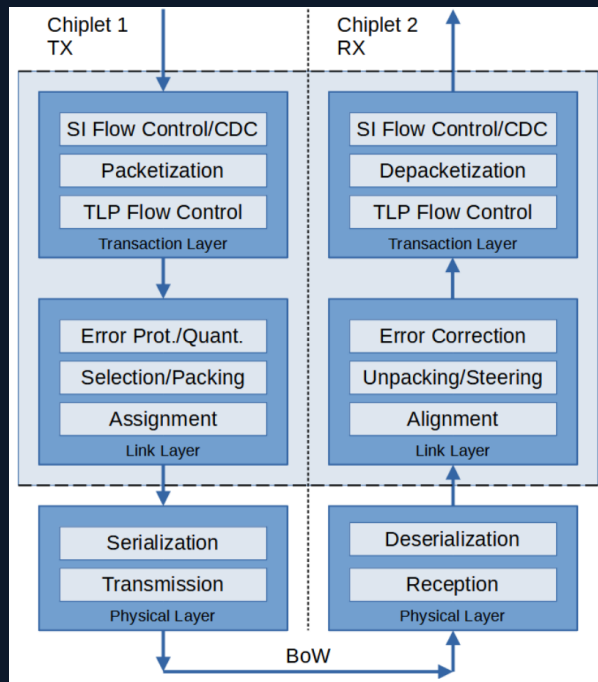
Common assumptions · Test vehicles · Owned interfaces

Original explanatory illustration · Sources: S03 pp12–14



A physical link is only one layer

CHAPTER 05 / CHIPLETS AND PROTOCOL LAYERS



Transaction → Meaning

Link → Protection and framing

Physical → Signals

Background 1.0 stack; the 2.0 changes come next

Brian Murray / VeriSilicon · “Advancing Chiplet Flexibility: Updates in 2.0 OCE Transaction and Link Layer Specification”
OCP Global Summit 2026 · PDF p3; background 1.0 diagram [S08]

What OCE TLL 2 changes

CHAPTER 05 / CHIPLETS AND PROTOCOL LAYERS



Format efficiency

PHY bindings

Defined profiles

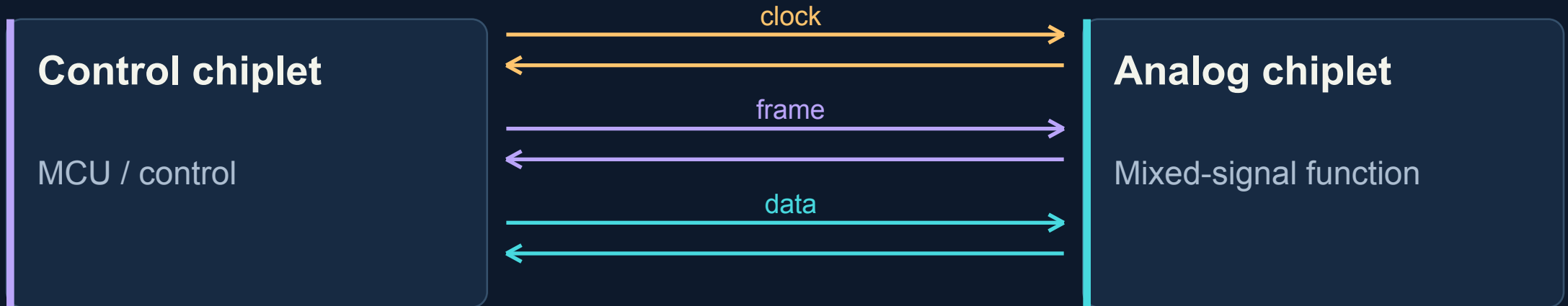
Packet size is a trade-off; compatibility with 1.x is limited

Original explanatory illustration · Sources: S08 pp5–10



Why a simpler interface still matters

CHAPTER 05 / CHIPLETS AND PROTOCOL LAYERS

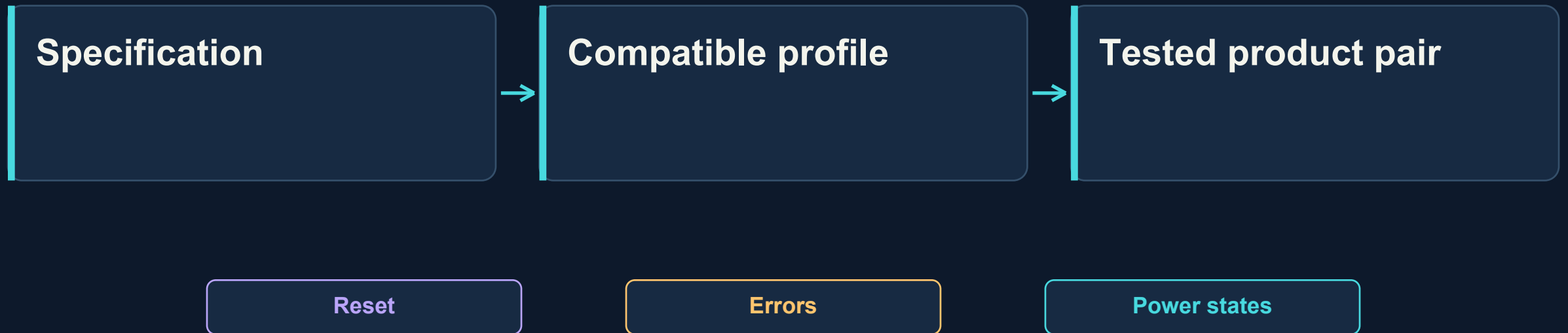


BoW.Flexi: single data rate · minimum six pins · configurable scope

Original explanatory illustration · Sources: S12 pp3–6

Interoperability is a product outcome

CHAPTER 05 / CHIPLETS AND PROTOCOL LAYERS

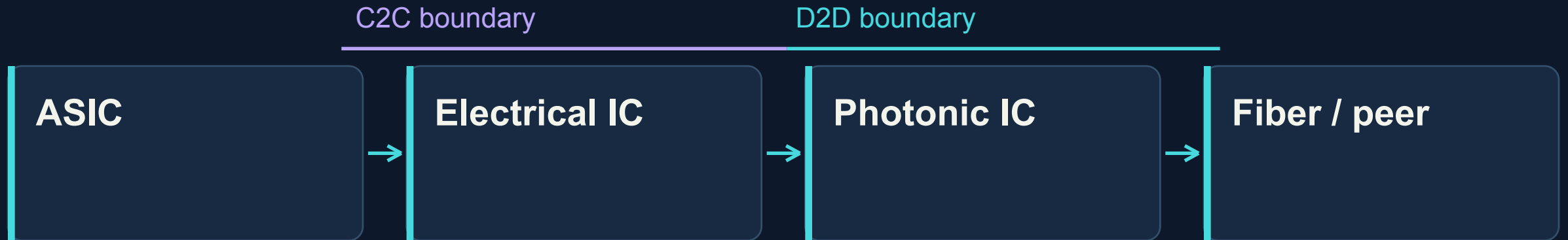


Which integration tasks disappear? Which remain?

Original explanatory illustration · Sources: S08 pp9–11 · S12 pp3–10

Follow the signal into the optical engine

CHAPTER 06 / OPTICAL INTERFACES AND OPERATIONS



ASIC: application-specific integrated circuit · IC: integrated circuit

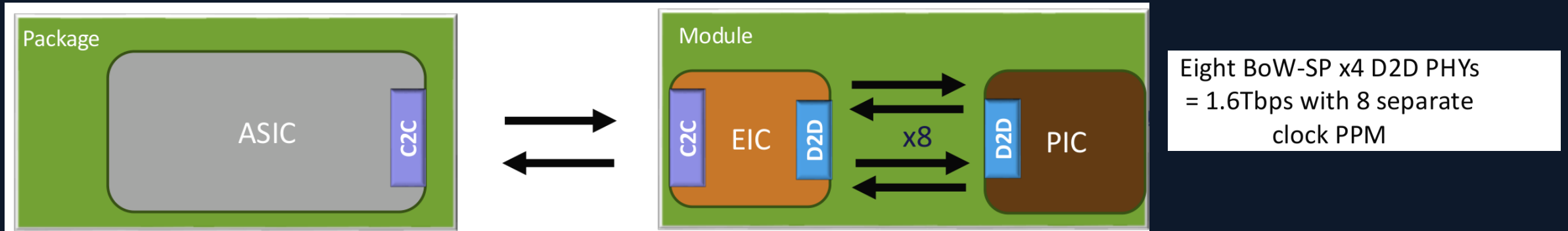
C2C: chip-to-chip · D2D: die-to-die · Cyan: data path

Electrical interfaces remain inside an optical system

Original explanatory illustration · Sources: S07 pp5–9

Fast and narrow or slow and wide

CHAPTER 06 / OPTICAL INTERFACES AND OPERATIONS



BoW-SP example, not BoW.Flexi
1.6T-class architecture; not application throughput

Channel organization determines lane and clock requirements

Bringing optics closer changes serviceability

CHAPTER 06 / OPTICAL INTERFACES AND OPERATIONS

Pluggable

Board-edge service
boundary

Near-packaged

Shorter electrical path
+ new integration

Co-packaged

Package integration
+ repair constraints

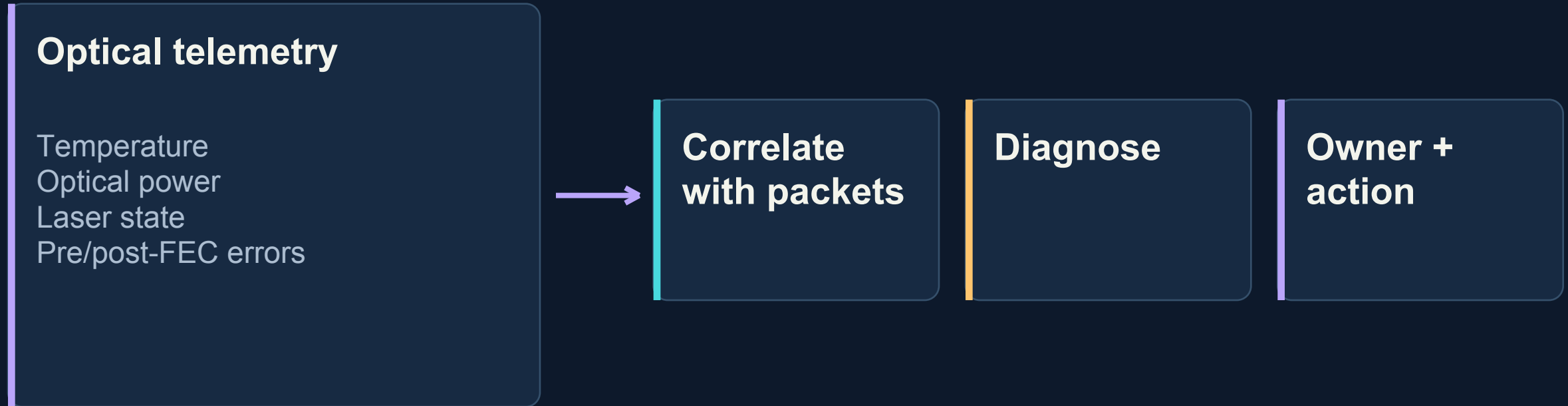
Evaluate the healthy path AND the repair path

Original explanatory illustration · Sources: S07 pp5–6,15–17 · S09 pp4–7



Optical health must be visible

CHAPTER 06 / OPTICAL INTERFACES AND OPERATIONS



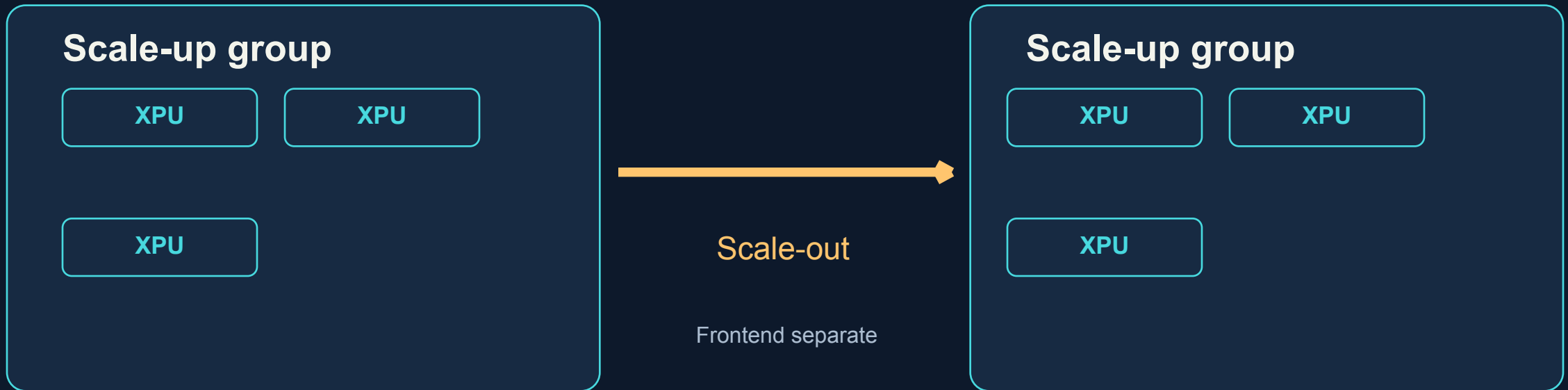
Correction can hide degrading margin

Original explanatory illustration · Sources: S09 pp7–8



Scale up and scale out

CHAPTER 07 / ETHERNET FOR SCALE-UP

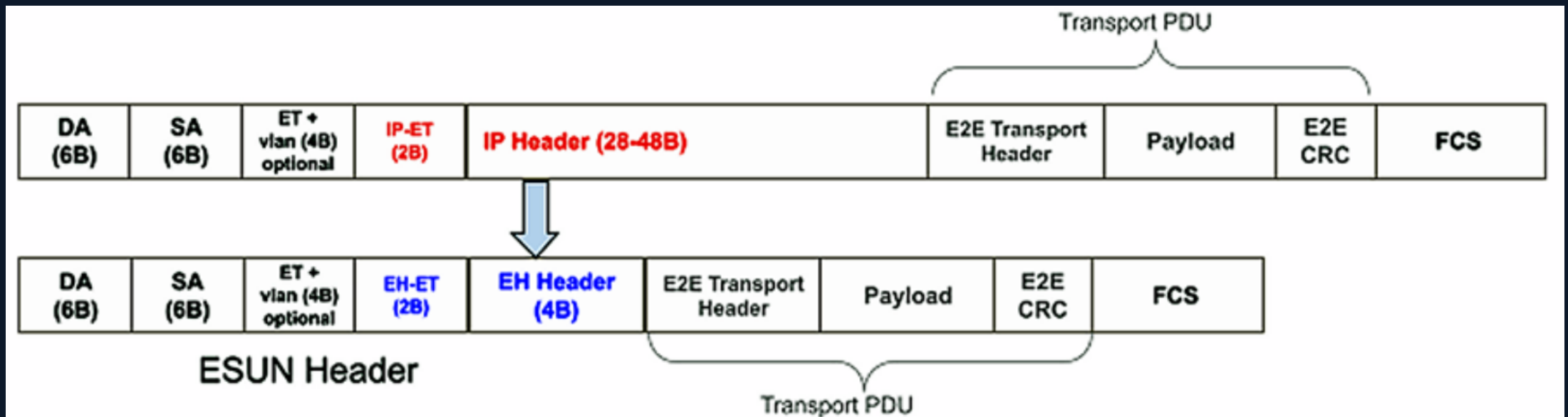


ESUN focuses on scale-up fabric requirements

Original explanatory illustration · Sources: S10 pp3–6

Small messages make overhead expensive

CHAPTER 07 / ETHERNET FOR SCALE-UP



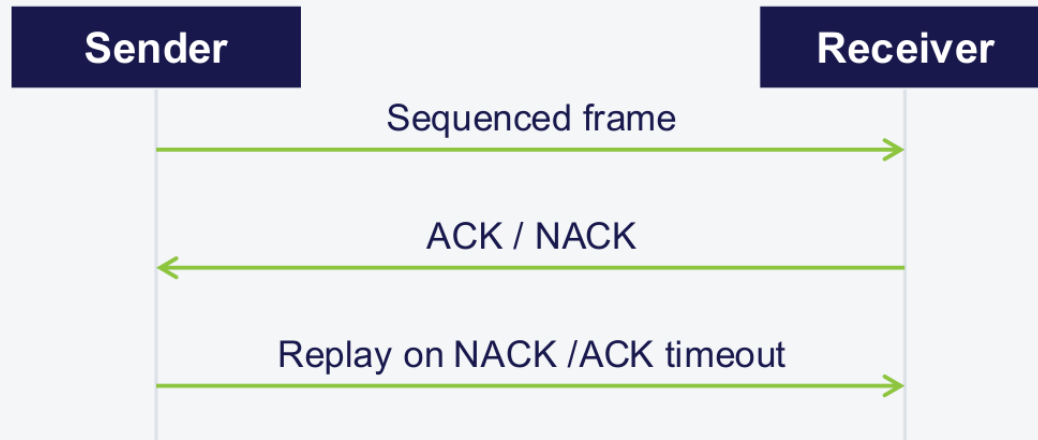
1.0 released · 1.1 optimizations were work in progress at the talk

Recover errors locally

CHAPTER 07 / ETHERNET FOR SCALE-UP

FOCUS: LINK-LEVEL RETRY

LLR — Link Level Retry



CBFC — Credit-Based Flow Control



Retry repairs transmission errors; recovery can add latency

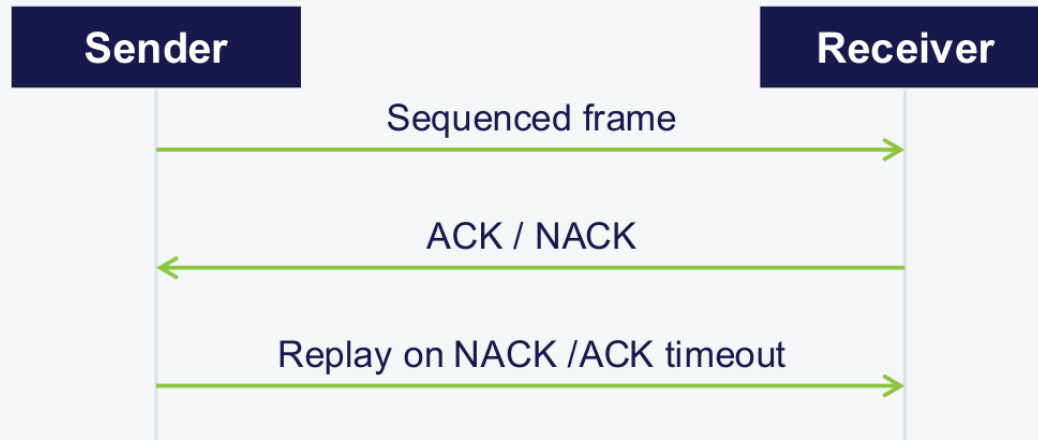


Credits prevent a different problem

CHAPTER 07 / ETHERNET FOR SCALE-UP

FOCUS: CREDIT-BASED FLOW CONTROL

LLR — Link Level Retry



CBFC — Credit-Based Flow Control



Credits protect receive buffers; round-trip time changes allowance



A passed test has a boundary

CHAPTER 08 / VALIDATION AND SYSTEM DECISIONS



Reported configurations and tests; not general certification

Original explanatory illustration · Sources: S11 pp7,10–15



Use one workload contract

CHAPTER 08 / VALIDATION AND SYSTEM DECISIONS

Memory
Software
Package

WORKLOAD CONTRACT

Model / precision
Context distribution
Concurrency / reuse
Output target + quality
Power boundary

Fabric
Operations

Keep each metric attached to its workload and boundary

Original explanatory illustration · Sources: S01 p9 · S02 p4 · S04 p12 · S06 pp20–21



The questions worth taking into a review

CHAPTER 08 / VALIDATION AND SYSTEM DECISIONS



Working set → path → integration → interoperability → failures → operations

Original explanatory illustration · Sources: S01–S12 · original leadership checklist



The architecture follows the data

CHAPTER 08 / VALIDATION AND SYSTEM DECISIONS



**Where? Reuse? Who moves it?
What boundary cost? What evidence?**

Twelve talks, one system view; full references accessible

Original explanatory illustration · Sources: S01–S12