STRATINV RESEARCH

TECHNOLOGY / AI INFRASTRUCTURE · ARCHITECTURE BRIEF

Disaggregated Inference on AWS

How AWS pairs **Trainium** for prefill with the **Cerebras CS-3 (WSE-3)** for decode, delivered through Amazon Bedrock

Prefill AWS Trainium handles it	**Decode** Cerebras CS-3 handles it	**Bedrock** Exclusive delivery surface	**Mar 2026** Collaboration announced

Prepared June 2026 · Sources: AWS & Cerebras announcements and technical commentary · For information purposes only

OVERVIEW

Executive summary

On **March 13, 2026**, Amazon Web Services and Cerebras Systems announced a collaboration to deliver what they describe as the fastest AI inference available in the cloud, accessed exclusively through **Amazon Bedrock**. The design rests on a technique called **inference disaggregation** — splitting a single inference request into two stages that have very different hardware demands, and running each on the silicon best suited to it.

AWS Trainium (Amazon's purpose-built AI chip) handles the **prefill** stage — ingesting and processing the user's prompt. The **Cerebras CS-3**, built on the third-generation **Wafer-Scale Engine (WSE-3)**, handles the **decode** stage — generating output tokens. The two are stitched together inside AWS data centers by high-speed **Elastic Fabric Adapter (EFA)** networking and presented to customers as a single logical service.

A note on terminology

This brief corrects two common mis-namings. The prefill chip is **AWS Trainium** (specifically Trainium3), **not** a Marvell product — there is no chip called "Triennium." Marvell is cited by analysts only as a potential beneficiary of the high-speed interconnect/IP around such hand-offs. The decode engine is the Cerebras **CS-3**, built on the **Wafer-Scale Engine (WSE-3)** — abbreviated **WSE**, not "WFE."

Why it matters

- **An architectural departure from GPU-centric inference.** It is the first time a major hyperscaler will host Cerebras wafer-scale hardware in its own data centers, and the first cloud availability of Cerebras' disaggregated inference solution.
- **Tuned for reasoning and agentic workloads.** Reasoning models generate many output tokens as they "think," which makes the decode stage the bottleneck — exactly where Cerebras' memory bandwidth is aimed.
- **Each chip does what it is best at.** Prefill is parallel and compute-heavy (Trainium's strength); decode is serial and memory-bandwidth-bound (the CS-3's strength). Splitting them lets each run at higher utilization.

THE ARCHITECTURE AT A GLANCE

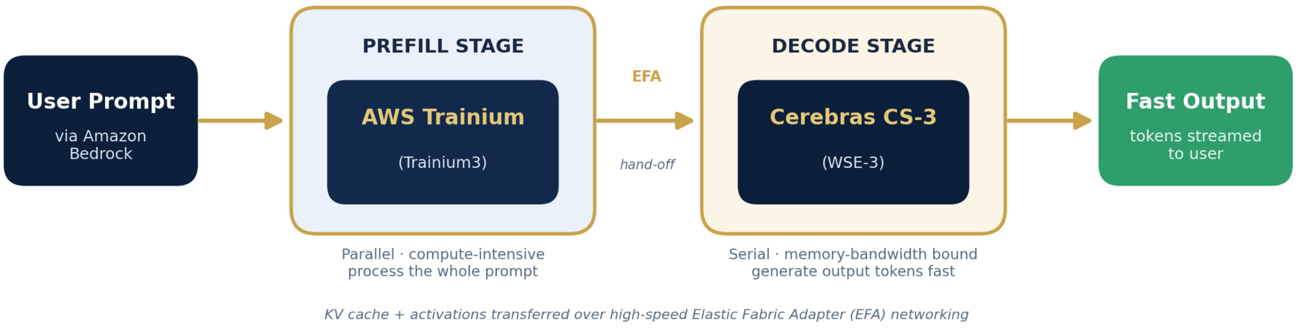

Figure 1 — Request flow: prompt enters via Bedrock, Trainium performs prefill, the state is handed to the Cerebras CS-3 over EFA for decode, and output tokens stream back.

THE CORE IDEA

Prefill and decode: two jobs, two chips

A transformer answering a prompt does two very different kinds of work. Disaggregation recognizes that a single accelerator optimized for one is necessarily compromised for the other — so AWS and Cerebras run them on separate, specialized pools of hardware.

PREFILL — AWS Trainium

Process the entire input prompt in one parallel pass, building the model's internal state (the KV cache) for every input token.

- Natively parallel — the whole sequence at once
- Computationally intensive (GEMM-bound)
- Requires only moderate memory bandwidth
- Maps well to Trainium's dense compute

DECODE — Cerebras CS-3 (WSE-3)

Generate output tokens one at a time, each step reading the full model + KV cache from memory. Throughput is gated by memory bandwidth, not raw FLOPs.

- Inherently serial — one token after another
- Computationally light per step
- Extremely memory-bandwidth intensive
- Maps to the WSE-3's on-wafer SRAM bandwidth

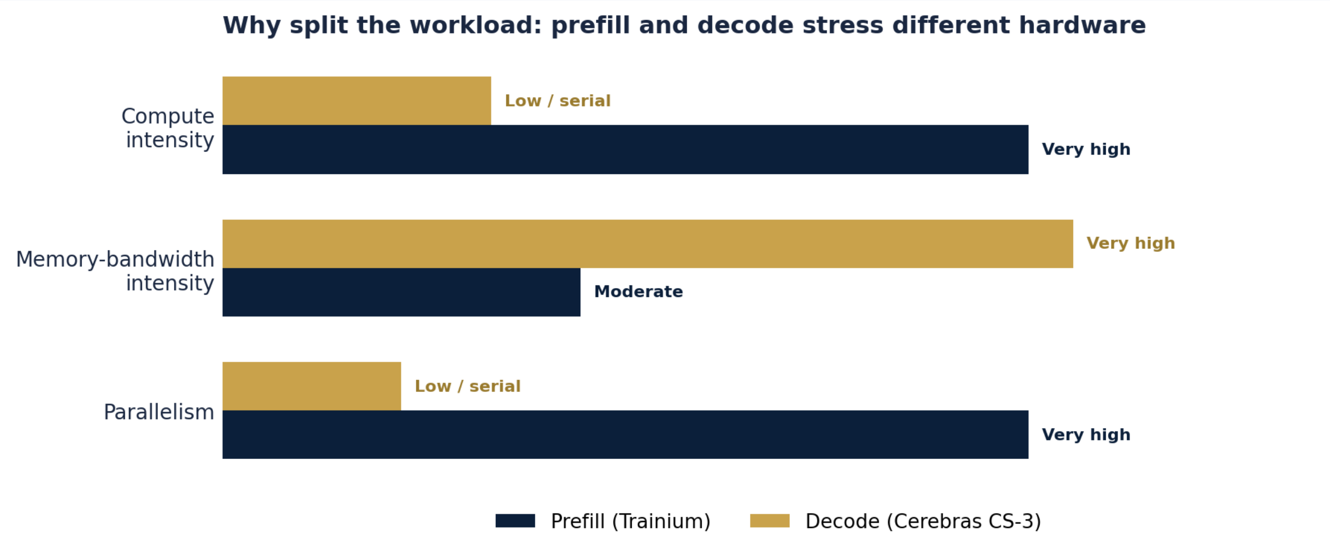

Figure 2 — Directional profile of the two stages (illustrative). Decode's memory-bandwidth demand is the property Cerebras' wafer-scale design targets most directly.

Why one chip can't ace both

Run everything on a GPU and the decode phase leaves much of the compute idle while starved for memory bandwidth; run everything on a bandwidth-optimized part and prefill under-uses that expensive bandwidth. By **disaggregating**, AWS can size each pool independently — scaling Trainium for prompt-heavy traffic and CS-3 capacity for long, token-heavy generations — so each processor delivers maximum tokens for its focused part of the workload.

THE BUILDING BLOCKS

What each component contributes

Component	Role in the system
AWS Trainium (Trainium3)	Amazon's purpose-built AI accelerator. Runs the prefill stage — parallel, compute-intensive prompt processing. Already used for training and inference of large models.
Cerebras CS-3 / WSE-3	Wafer-scale engine described as the world's fastest AI inference system, with on-wafer memory bandwidth Cerebras cites as thousands of times a leading GPU's. Dedicated entirely to decode.
Elastic Fabric Adapter (EFA)	High-bandwidth, low-latency AWS networking that carries the hand-off (KV cache / activations) between the prefill and decode pools with RDMA-class performance.
AWS Nitro System	Provides the secure isolation and virtualization that lets heterogeneous silicon operate as one trusted service inside AWS.
Amazon Bedrock	The managed model-serving surface where the combined solution is exposed. AWS is the first — and exclusive — cloud for Cerebras' disaggregated inference.

The request lifecycle, step by step

1

Request arrives
A user/application calls a model through Amazon Bedrock.

2

Prefill on Trainium
Trainium ingests the full prompt in parallel and builds the KV cache.

3

Hand-off over EFA
The prefill state is transferred to the decode pool via low-latency EFA networking.

4

Decode on CS-3
The Cerebras CS-3 generates output tokens serially at high speed, bandwidth-bound.

5

Stream back
Tokens stream to the caller through Bedrock as a single, seamless response.

READING THE MOVE

Strategic implications

Breathes new life into Trainium

By giving Trainium a clearly defined, high-value role (prefill) inside a best-of-breed pipeline, AWS strengthens the case for its own silicon rather than positioning it as a straight GPU substitute.

A credibility anchor for Cerebras

First hyperscaler to host Cerebras wafer-scale hardware in its own data centers; pairs Cerebras' speed story with AWS' distribution and enterprise reach via Bedrock.

Pressure on GPU-centric inference

Demonstrates a production path where specialized parts beat a single general-purpose accelerator on speed for reasoning/agentic workloads — a competitive signal to incumbents.

Ecosystem beneficiaries

Analysts note that high-speed hand-off between pools rewards interconnect/IP suppliers (e.g., Marvell, Synopsys) — the most defensible link between this architecture and Marvell.

Caveats & open questions

- **Status & timing.** Announced March 2026 as rolling out "in the coming months" via Bedrock; treat exact availability, pricing and benchmarks as not-yet-fully-public and subject to change.
- **Hand-off complexity.** Coordinating a Trainium prefill cluster with a Cerebras decode cluster requires sophisticated load-balancing and latency management; real-world efficiency depends on that orchestration.
- **Benchmarks pending.** Headline speed claims are vendor-stated. Independent latency/throughput and cost-per-token comparisons against GPU-based inference will be the real test.
- **"Triennium" / "WFE" do not exist as named products.** Any analysis should be built on Trainium and the Cerebras CS-3 / WSE-3 to avoid propagating mis-specified architecture claims.

Bottom line

AWS is not using a "Marvell Triennium." It is pairing **its own Trainium** (prefill) with the **Cerebras CS-3 / WSE-3** (decode) over EFA, served through Bedrock — a disaggregated design that puts each workload on the hardware built for it. The concept is sound and well-documented; the open variables are availability timing, orchestration efficiency, and independent benchmarks.

Sources: AWS & Cerebras joint announcement and press materials (Mar 13, 2026); coverage and technical commentary from Cerebras, DataCenterDynamics, Futurum Group, HPCwire/AIwire, Techzine and others (Mar–Jun 2026). Prepared by Stratinv Research, June 2026. For information purposes only — not investment advice.