

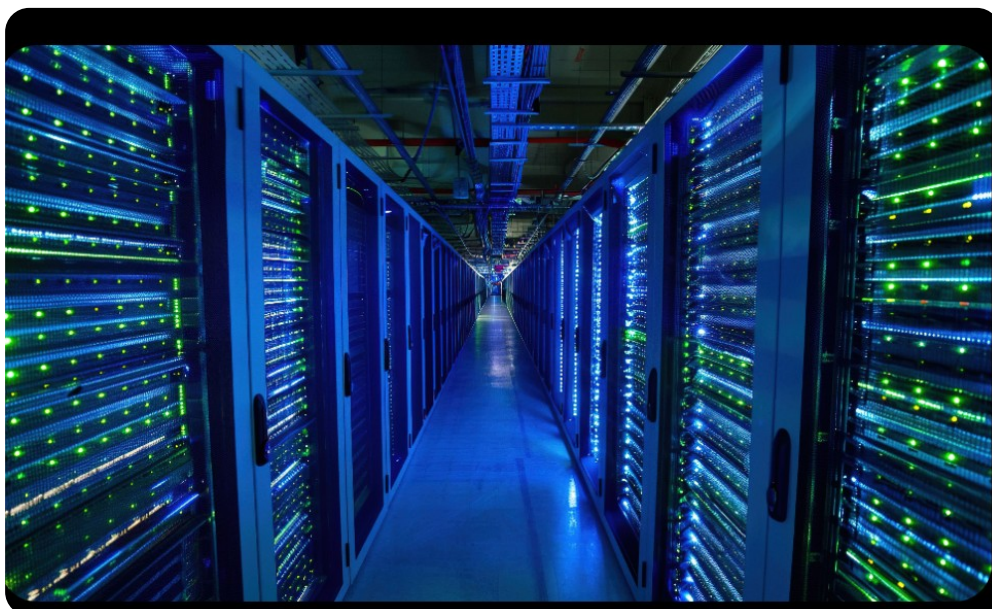
Blogue Fragmentos do Caos



A verdade nasce onde o pensamento é livre.

DGX Station + DeepSeek: engenharia de IA local com memória coerente, contexto longo e agentes com ferramentas

Publicado em 2026-03-02 11:06:27



BOX DE FACTOS

- **DGX Station:** primeira workstation “deskside” com o **NVIDIA GB300 Grace Blackwell Ultra Desktop Superchip**, com **~784GB** de memória unificada/coerente (a NVIDIA refere também ~775GB em

Blogue Fragmentos do Caos



A verdade nasce onde o pensamento é livre.

pensado para multi-station scaling e integração com infra-estrutura empresarial.

- **DeepSeek-V3.2** (01/12/2025): primeiro modelo da DeepSeek a integrar “thinking” directamente no **tool-use** e a suportar ferramentas em modo “thinking” e “non-thinking”.
- **DeepSeek-V3.2-Exp** (29/09/2025): introduz **DeepSeek Sparse Attention (DSA)** para contexto longo com menor custo computacional; anunciado como passo intermédio para a próxima arquitectura.
- **Mensagem técnica:** a combinação “memória coerente massiva + contexto longo eficiente + tool-use” torna viável a IA local para RAG, agentes e automação empresarial com dados sensíveis “em casa”.

Blogue Fragmentos do Caos



A verdade nasce onde o pensamento é livre.

contexto longo e agentes com ferramentas

A era do “modelo na nuvem” continua — mas há um regresso silencioso ao terreno: executar, governar e auditar IA dentro da própria infra-estrutura. A DGX Station é o ferro; o DeepSeek V3.2 é a mente que aprende a operar com ferramentas.

Imagem — (substituir pelo URL depois de carregares no WordPress)



Blogue Fragmentos do Caos



A verdade nasce onde o pensamento é livre.

uma máquina concebida como **plataforma coerente**: CPU+GPU ligados por interconexão de altíssima largura de banda, com **memória unificada/coerente em escala anormal** para o mundo “deskside”. Isto altera o tipo de trabalho que se consegue fazer localmente: modelos maiores, contextos longos, múltiplos serviços, e pipelines completos (indexação, inferência, pós-processamento, auditoria) sem “trocas de contexto” de infra-estrutura.

A NVIDIA posiciona a DGX Station como o primeiro sistema “deskside” com o **GB300 Grace Blackwell Ultra**, com **~784GB** de memória unificada e conectividade de rede até **800Gb/s** via **ConnectX-8**. O recado é claro: não é apenas potência; é **capacidade de caber**. E “caber” é o que decide se um modelo e o seu contexto vivem localmente ou se são expulsos para a nuvem.

2) Memória coerente: a arma silenciosa

Em IA aplicada, o gargalo real raramente é “FLOPS”. É **memória**: quanto cabe, com que latência, e com que custo

Blogue Fragmentos do Caos



A verdade nasce onde o pensamento é livre.

modelo) sem amputações brutais;

- **Contexto longo** sem “pagamento” excessivo em trocas CPU↔GPU;
- **RAG robusto** com indexação, re-ranking e caches persistentes no mesmo nó;
- **Serviços paralelos** (API, embeddings, vector DB, auditoria, observabilidade) sem a máquina entrar em asfixia.

Em termos de arquitectura, isto dá-te liberdade: podes desenhar um “nó IA” que seja simultaneamente laboratório e produção, com camadas de governação e logging que, em cloud, muitas vezes ficam subfinanciadas por custo.

3) DeepSeek V3.2: “thinking” a operar com ferramentas

O salto do DeepSeek-V3.2 é conceptual e prático: integrar “thinking” no **tool-use**. Isto aproxima o modelo do papel de **agente** (planeia, chama ferramentas, valida, volta a chamar, conclui) em vez de ser apenas um “oráculo textual”. A própria DeepSeek refere treino de agentes com síntese massiva de dados cobrindo milhares de ambientes e dezenas de milhares de instruções complexas.

Blogue Fragmentos do Caos



A verdade nasce onde o pensamento é livre.

ticket, produzir relatório. É aqui que a IA deixa de ser conversa e passa a ser **processo**.

4) V3.2-Exp e DSA: contexto longo com menos imposto computacional

A DeepSeek-V3.2-Exp introduz **DeepSeek Sparse Attention (DSA)**, apontada como mecanismo para reduzir custo computacional e melhorar desempenho em texto de sequência longa. Isto interessa porque a IA empresarial vive de **documentos longos**: procedimentos, contratos, catálogos, histórico de suporte, relatórios técnicos, normativos. Contexto longo é a diferença entre “resposta bonita” e “resposta correcta, com fonte e coerência”.

5) Arquitectura recomendada: DGX Station como nó on-prem (padrão de produção)

Um desenho “limpo” (e auditável) para pôr DeepSeek em produção na DGX Station:

Blogue Fragmentos do Caos



A verdade nasce onde o pensamento é livre.

datas).

2. **Embeddings**: geração de vectores e armazenamento em **vector DB** (com colecções por domínio e ACLs).
3. **RAG**: retrieval + re-ranking + montagem de contexto com citações internas (IDs de documento e páginas).
4. **Inferência**: DeepSeek-V3.2 como “motor”, com **tool-use** activado e limites de segurança.
5. **Ferramentas**: conectores (ERP/CRM/SQL) e funções de negócio (ex.: cálculo de orçamento, validação de margens).
6. **Validação**: regras determinísticas + verificação de campos críticos (ex.: valores, impostos, condições).
7. **Auditoria**: logging de prompts, fontes, chamadas a ferramentas, tempos, erros e versões de modelo.

O benefício de executar isto num nó “deskside” potente é a **latência baixa** e a **soberania de dados**. O benefício de ter “tool-use” é transformar respostas em **acções controladas**. E o benefício de DSA/contexto longo é reduzir o risco de respostas fora de contexto quando os documentos crescem.

Blogue Fragmentos do Caos



A verdade nasce onde o pensamento é livre.

compensa usar quantização (quando compatível) para ganhar throughput e reduzir pegada, mantendo qualidade aceitável.

- **Separação de serviços:** mesmo num nó único, isola “API gateway”, “RAG service”, “model server” e “tools service” para observabilidade e controlo de falhas.
- **Cache e warmup:** caches de embeddings e de re-ranking (por colecção) reduzem custos e melhoram previsibilidade.
- **Políticas:** define claramente o que o agente pode fazer. “Tool-use” sem guarda é automação cega; com guarda, é produtividade.

Epílogo: 2026 é o ano em que a IA volta a ser infra-estrutura

A DGX Station representa uma ideia antiga com roupa nova: **computação perto dos dados**. O DeepSeek V3.2 representa uma ideia moderna com músculo: **modelos que raciocinam enquanto operam ferramentas**. Juntos, dão forma a uma arquitectura onde a IA não é um

Blogue Fragmentos do Caos



A verdade nasce onde o pensamento é livre.

- NVIDIA — DGX Station (página oficial): <https://www.nvidia.com/en-eu/products/workstations/dgx-station/>
- NVIDIA Newsroom (18 Maio 2025) — DGX Station: GB300, 20 PFLOPS, 784GB, ConnectX-8 800Gb/s: <https://nvidianews.nvidia.com/news/nvidia-launches-ai-first-dgx-personal-computing-systems-with-global-computer-makers>
- DeepSeek — V3.2 Release (01 Dez 2025): <https://api-docs.deepseek.com/news/news251201>
- DeepSeek — Introducing V3.2-Exp (29 Set 2025): <https://api-docs.deepseek.com/news/news250929>
- Reuters (29 Set 2025) — V3.2-Exp, DSA, “intermediate step”: <https://www.reuters.com/technology/deepseek-releases-model-it-calls-intermediate-step-towards-next-generation-2025-09-29/>

Francisco Gonçalves

Fragmentos do Caos • FC-Chronic-News

Co-autoria editorial: Augustus



Blogue Fragmentos do Caos

A verdade nasce onde o pensamento é livre.



Fragmentos do Caos:

[Blogue](#)

•

[Ebooks](#)

•

[Carrossel](#)



Esta página foi visitada ... vezes.

[Contactos](#)