Systems Innovation for Agentic AI

The DASE Architecture and the VAST AI OS

Dr. James Chen
Field Technical Director - HPC, APJ
james.chen@vastdata.com

Agenda

- Vast Data
- DASE Architecture & VAST AI OS
- Breaking the Memory Wall for Scale-out Inference
- VAST Foundation Stacks with NVIDIA AI Blueprints

OPEN Compute Project®

开放计算标准工作委员会 Open Compute Technology Committee

VAST

The World's Leading AI Clouds Are Powered By VAST

Photo: First Cloud-Provider GB200 Bring-Up • The CoreWeave Stack Runs on VAST

CoreWeave

2026 VAST Data Overview

Building The Next Generational Data Infrastructure Company

FOUNDED

2016

10 Years Old

TEAM

1,200+

Worldwide VASTronauts

FUNDING

Cash Flow Positive

Fueled By Our Business

HQ

GTM: NYC
R&D: TLV

Operating in Dozens of Countries

Industry Leaders Choose VAST

Leading AI Clouds

- CoreWeave
- Lambda
- Scaleway
- core42 (A G42 company)
- SK telecom
- together.ai
- NSCALE

Leading Financial Services

- Man
- jumptrading
- G RESEARCH
- CA
- AQUATIC
- HSBC
- SQUARE POINT

Technology Leaders

- servicenow
- xAI
- ALEPH ALPHA
- DigitalOcean
- zoom
- agoda
- LG

Leading Health Orgs

- CHAN ZUCKERBERG Biohub Network
- labcorp | INVITAE
- 総合病院 国保旭中央病院
- Boston Children's Hospital (Until every child is well)
- Oxford Nanopore Technologies
- EMBL
- NIH

Leading HPC Centers

- Lawrence Livermore National Laboratory
- HARVARD UNIVERSITY
- CINECA
- NERSC
- NIWA (Taihoro Nukurangi)
- TACC (Texas Advanced Computing Center)
- CSCS (Centro Svizzero di Calcolo Scientifico, Swiss National Supercomputing Centre)

Leading Media & Telcos

- PIXAR
- Virgin mobile
- NHL
- FRAMESTORE
- MONUMENTAL
- STREAMLAND MEDIA
- COMPANY3

The Rise of AI Inference
Everything Changes

FRONTIER MODEL TRAINING	INFERENCE ERA REQUIREMENTS
Batch Pipelines	**Real-Time Pipelines** Unpredictable & Always-On
Highly Curated Datasets	**AI Enterprise Data** Accessed Non-Deterministically
Single Tenant AI Labs	**Multi-Tenant Services** Strict Data Isolations
Days-Months: Just Training	**A Continuous AI Flywheel** Inference & Find Tuning

VAST Data's DASE Architecture

Disaggregated, Shared Everything

Linear Scalability

100,000s of CPUs/GPUs, Exabytes of Data

Real-Time At Scale

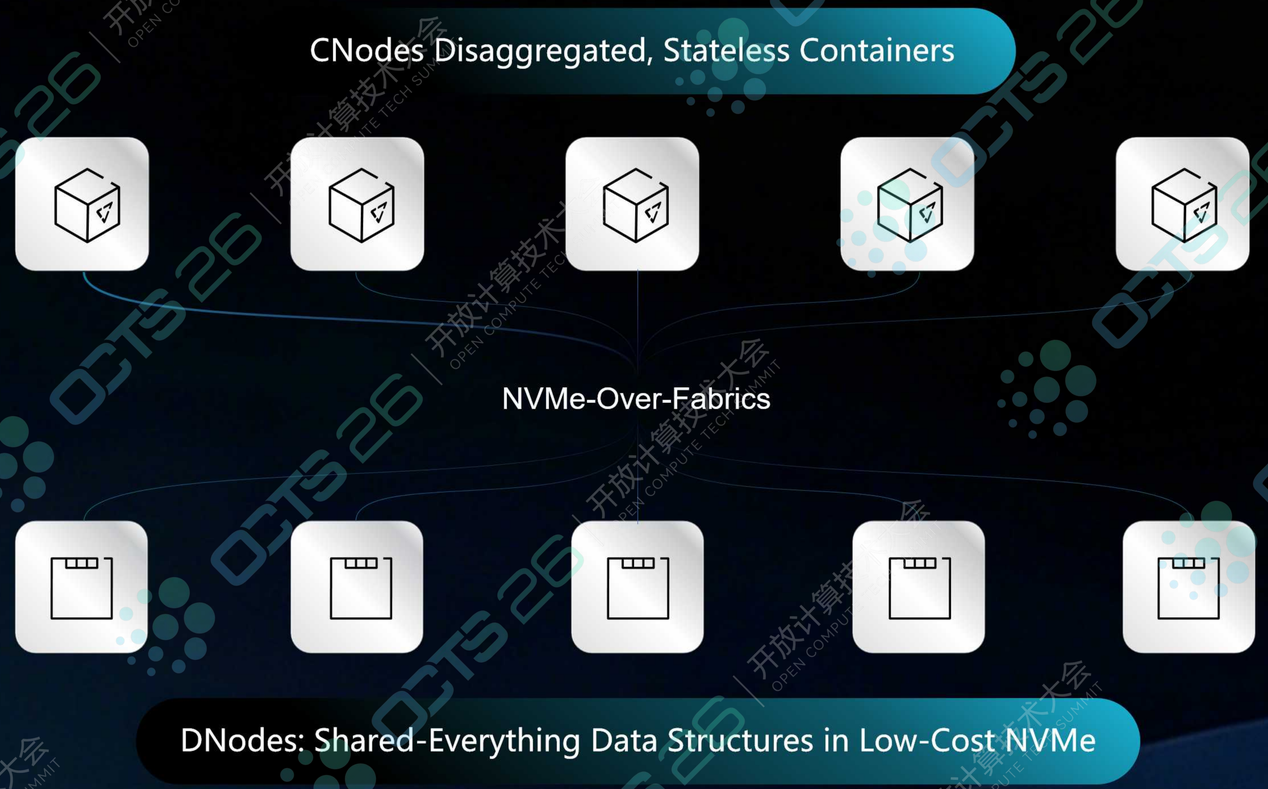
Parallel File I/O, Events, Vectors, Queriest

Resilient at Extreme Scale

99.999+% Uptime In The World's Harshest DCs

Radically Affordable

Accelerate CPUs and Lower SSD cost



The VAST AI Operating System

DataEngine
Scalable, Event-Driven Computing

Triggers, Functions, Containers

SyncEngine
Index & DataRouter

Massive-Scale

InsightEngine
Real-Time RAG

Embedding Creation

AgentEngine
Tools & Orchestration

Agents, Tools, Observability

DataStore
Universal Storage Infrastructure

NFS, SMB, S3, NV/Me/TCP

DataBase
Transactional & Analytical Database

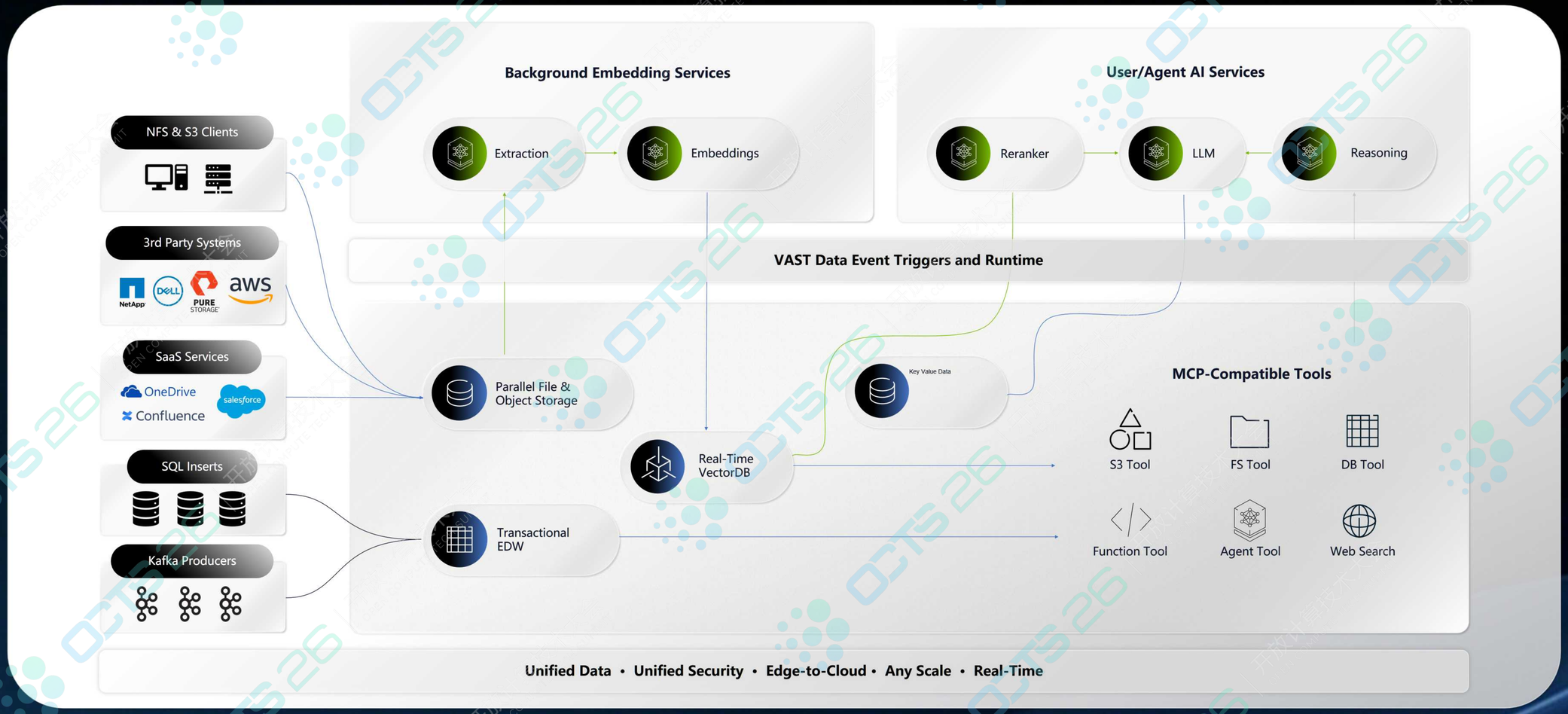
Kafka, Python, Parquet, SQL, Similarity Vectors

DataSpace
Globally-Distributed Data Computing

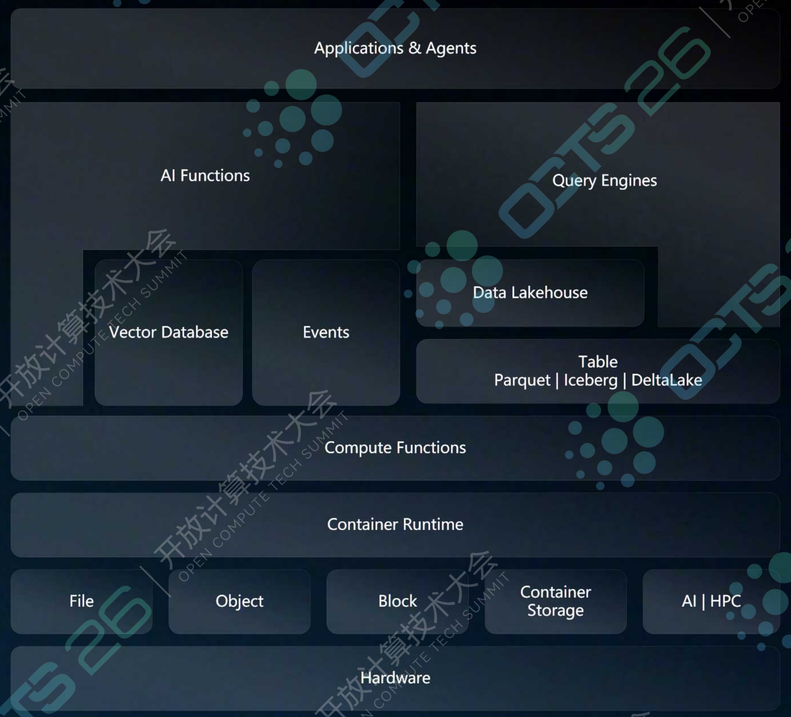
Hardware Platforms: Cisco, Dell Technologies, Hewlett Packard Enterprise, Supermicro

Traditional CSPs: aws, Google Cloud, Oracle

GPU Clouds: Equinix, CoreWeave, Core42, Lambda, and more..


Background Embedding Services
Extraction
Embeddings
User/Agent AI Services
Reranker
LLM
Reasoning
NFS & S3 Clients
3rd Party Systems
NetApp
DELL
PURE STORAGE
aws
SaaS Services
OneDrive
Confluence
salesforce
SQL Inserts
Kafka Producers
VAST Data Event Triggers and Runtime
Parallel File & Object Storage
Key Value Data
Real-Time VectorDB
Transactional EDW
MCP-Compatible Tools
S3 Tool
FS Tool
DB Tool
Function Tool
Agent Tool
Web Search
Unified Data • Unified Security • Edge-to-Cloud • Any Scale • Real-Time

Much More than Storage – VAST Delivers a Unified AI Data Platform



Fragmented AI Infrastructure

10+ products to audit, tune, integrate, maintain, govern, and scale.
Operational inefficiency, multiple data copies,
High management overhead and siloed systems

SPOC for Support

Unified AI Platform

Reduces $$$, complexity, silos and management overhead
Simplified governance, reduced operational burden
Unified security and performance at scale

Agentic AI Redefines Inference

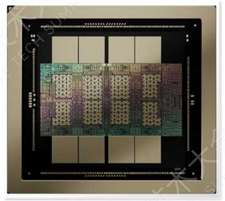
Multi-turn, reasoning-based AI introduces new inference infrastructure demands

- AI shifts from chatbots to multi-step workflows
- Agents require persistent inference context
- Context grows with longer sequences and sessions
- KV cache becomes long-lived AI memory
- Performance depends on storing and reusing context

Inference Context is the New Bottleneck

Storage must be re-architected

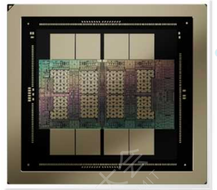
- Inference context is large, dynamic, and re-computable
- Context must be shared across GPUs and nodes
- Local memory is inherently limited
- Scaling traditional storage increases cost and power
- Efficient context scaling demands a new architecture



Context Memory Hierarchy

GPU HBM	G1 Active KV
System Memory	G2 Staging / Spillover KV
Local Storage	G3 Warm KV reuse
Network Storage	G4 Cold or shared KV context

KV Cache Offload and Acceleration with VAST Data Services

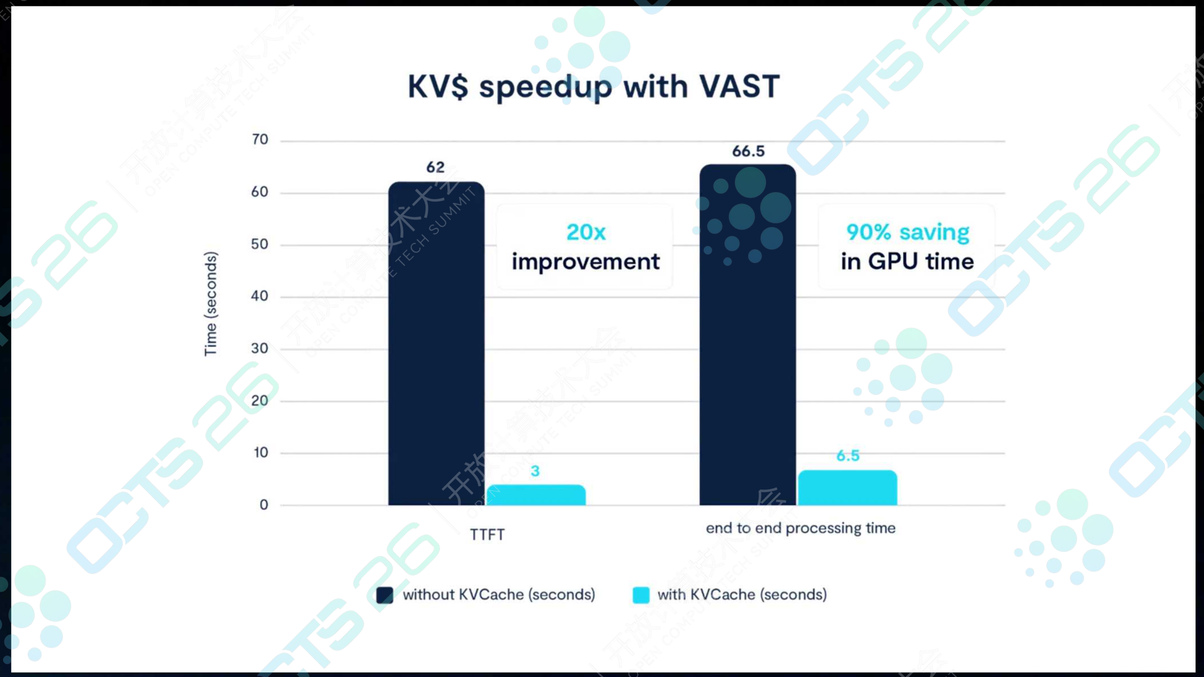


GPU HBM	G1 Active KV		
System Memory	G2 Staging / Spillover KV		
Local Storage	G3 Warm KV reuse	NFS RDMA/TCP	VAST **DataStore** • Hyperscale File & Object Storage • World-Leading Data Reduction • Enterprise Data Services + AI Scalability
Network Storage	G4 Cold or shared KV context	S3 RDMA/TCP	

Context Memory Hierarchy

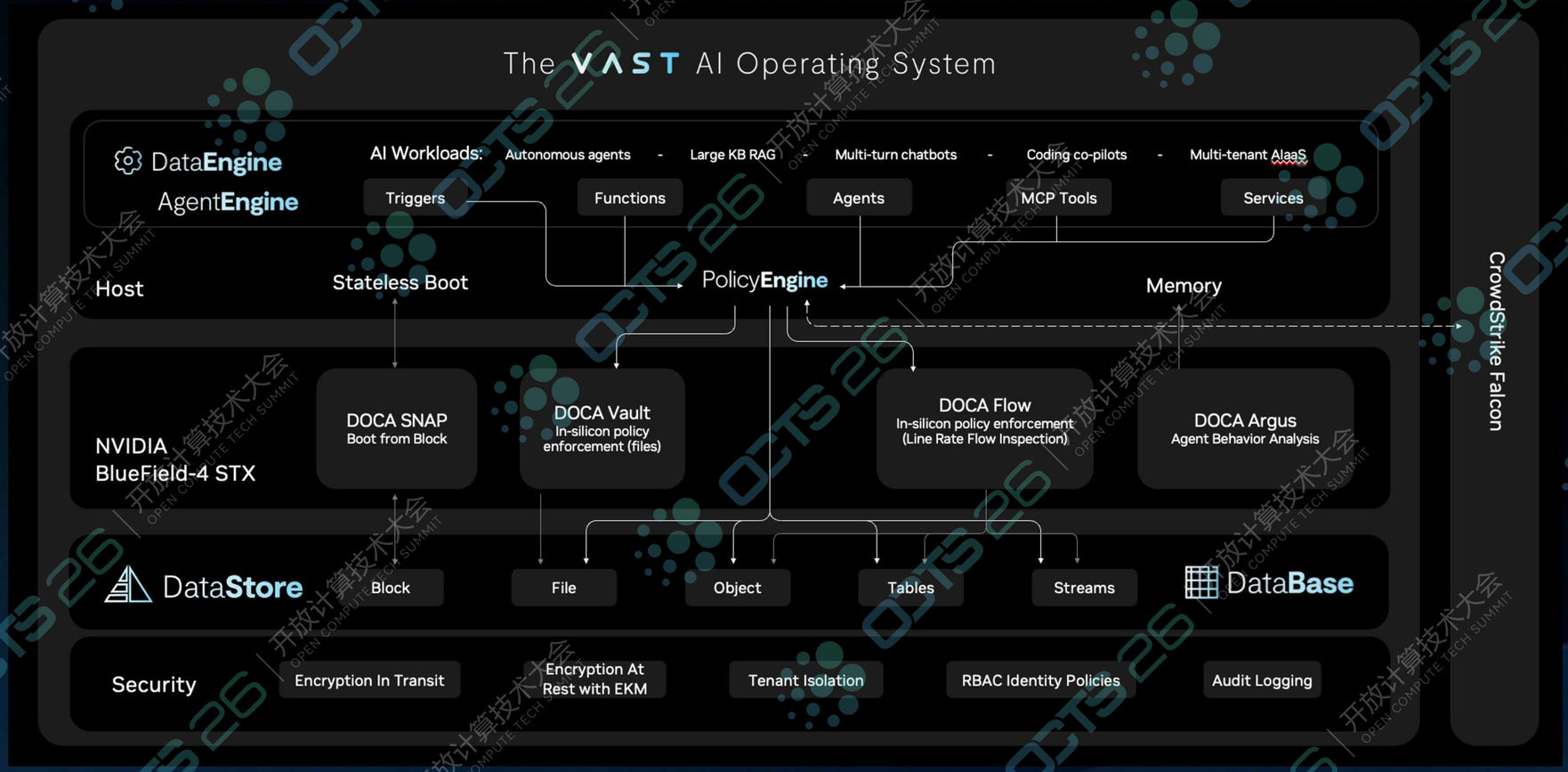
Tests shows highly performant storage access, with potential for better results with increased network bandwidth

Blog: https://www.vastdata.com/blog/nvidia-dynamo-vast-scalable-optimized-inference



NVIDIA HGX™ H100 (8*H100-80GB HBM3)
NVIDIA Dynamo pre-release 0.7.0 (Nov 21, 2025, modified) + vLLM v0.11.0 (modified) + NIXL 0.7.1
2 x 100 Gbit/s network bandwidth
clowman/Llama-3.1-405B-Instruct-AWQ-Int4

End-to-End Secure AI with VAST Data, NVIDIA Vera Blue Field-4, and NVIDIA DOCA



Blog: https://www.vastdata.com/blog/vast-zero-trust-agentic-ai-nvidia-bluefield-doca

AI-Ready, Fine-Grained Data Governance & Access Control

Unified AI Governance
Ensure consistent security from raw data to vectors for unified governance across AI pipeline

Row & Column Permissions
Control access at the most granular level across VAST DataBase

Regulatory Compliance & Risk Mitigation
Prevent unauthorized access, enable comprehensive observability, and ensure data privacy

VAST Foundation Stacks Turn NVIDIA AI Blueprints into Adoption and Enterprise Ready Deployments

- **Extend NVIDIA AI Blueprints**
  with repeatable, adoption-ready implementations designed to run natively on the VAST AI Operating System.
- **Unify**
  data access, database services, compute orchestration, eventing, and pipeline execution in a single environment.
- **Deploy**
  scalable AI pipelines without building complex infrastructure from scratch
- **Seamless and repeatable deployment**
  anywhere the VAST AI OS runs, on-prem or in the cloud

VAST Foundation Stack for Video Search and Summarization

- **Get real-time insight from archive or streaming video**
  Ingestion pipeline with event triggers automates video insight embeddings
- **Secure query access with enterprise authentication**
  High performance vector DB with built-in access control ensures fast, secure and accurate query results.
- **Powerful combination of NVIDIA VSS blueprint with VAST Data Platform**
  Fast track VSS use cases with adoption-ready code that can be customized and deployed quickly

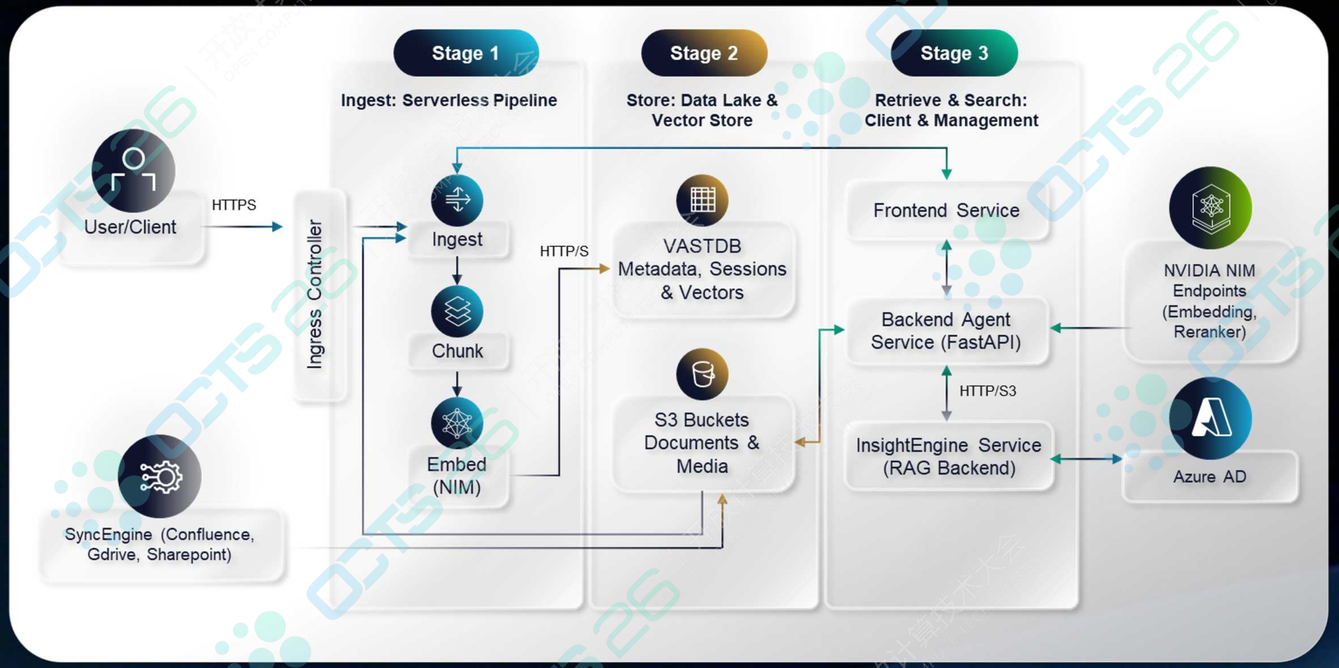
Real-Time Video Search and Analysis powered by VAST DataEngine, using NVIDIA NIM's & Cosmos-Reason Models

https://github.com/vast-data/cosmos-labs/tree/main/dataengine-vss-blueprint

VAST Foundation Stack Research Assistant Agent (RAG)

- **RAG-Powered Q&A**
  Retrieval Augmented Generation on document collections using VAST DataEngine and InsightEngine
- **Flexible Model Support**
  Works with local NVIDIA GPU endpoints or remote NVIDIA NIM APIs for LLM, embedding, and reranker models
- **Enterprise Authentication**
  Optional Azure AD integration for identity-based access control to document collections

AI Agents connect, retrieve and reason on an enterprise unified data platform

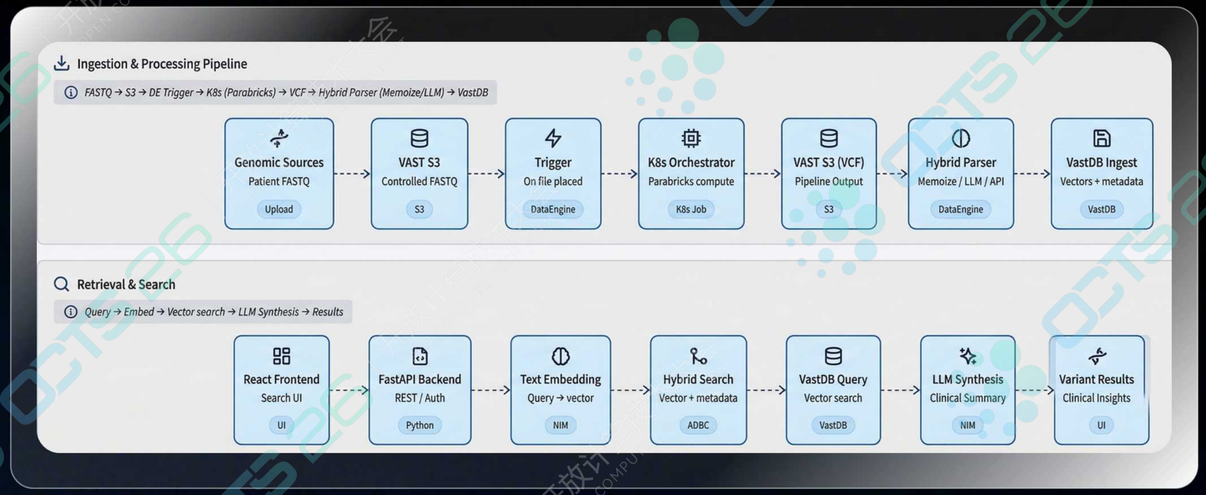


https://github.com/vast-data/cosmos-labs/tree/main/dataengine-research-assistant-blueprint

VAST Foundation Stack
Genomic RAG Engine

- **Life Sciences and Health Care RAG Engine** - turns DNA sequencing into AI-powered clinical insights on a single, unified platform
- **Accelerate Patient Treatment** - from months to hours with automated DNA analysis, research annotation, and recommendations for drug discovery with NVIDIA BioNeMo
- **RAG system transforms raw genomic data** into actionable, ranked clinical insights by with VastDB vector search with LLM synthesis
- **End-to-end Genomic Pipeline and Clinical** Search system powered by VAST DataEngine, NVIDIA Parabricks, and NVIDIA NIMs

Accelerate DNA to Precision Medicine and Drug Discovery from Months to Hours



https://github.com/vast-data/genomic-engine

THANK YOU

Dr. James Chen
Field Technical Director - HPC, APJ
james.chen@vastdata.com
www.linkedin.com/in/jenchangchen