AKKA

QCon London

A Blueprint for Agentic AI Services

Best practices for designing and operating agentic-scale services

Accelerate delivery. Stay safe and be efficient.

Today’s discussion

AKKA

01 Welcome

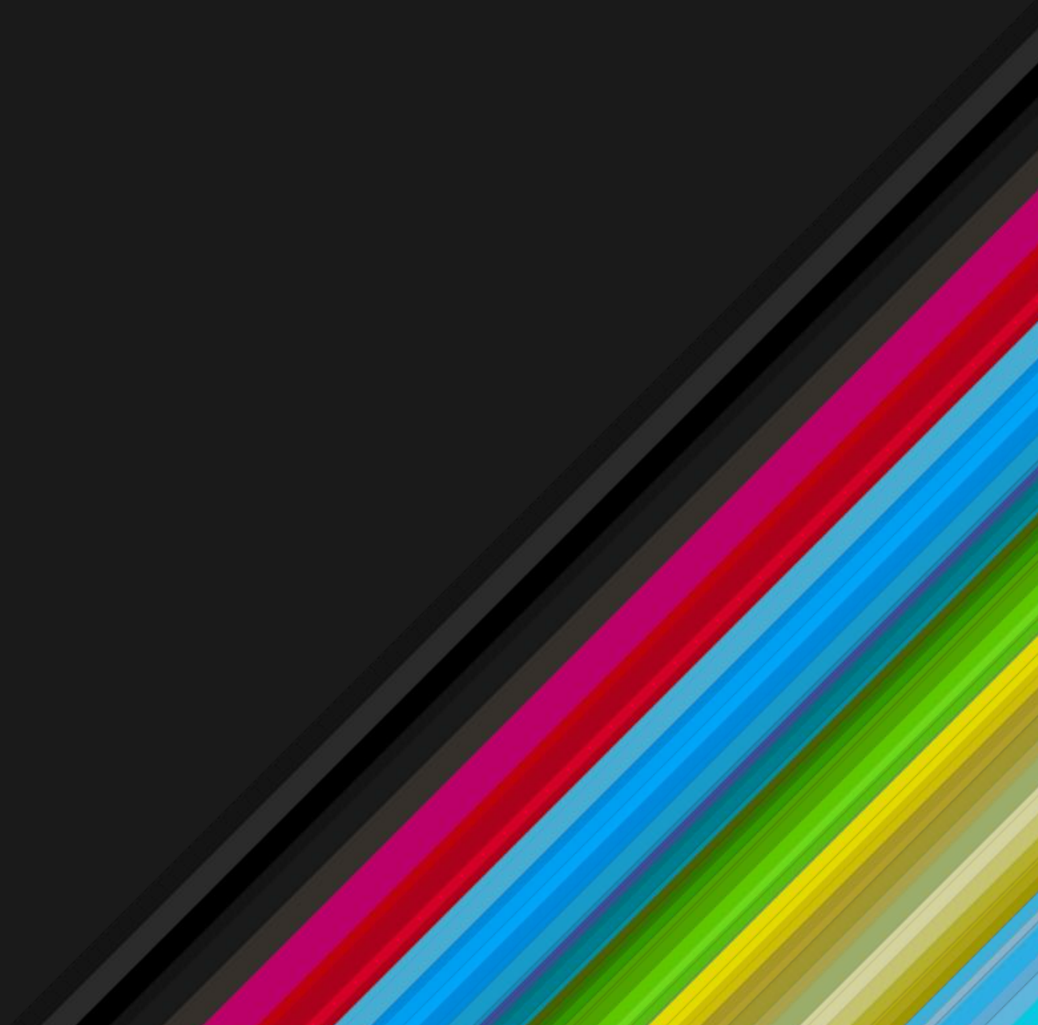
Duncan DeVore, Sr. Director, Architect & AI Advocate, Akka

02 What is Agentic AI?

03 The challenges of Agentic AI

04 A blueprint for Agentic Services

05 Q&A



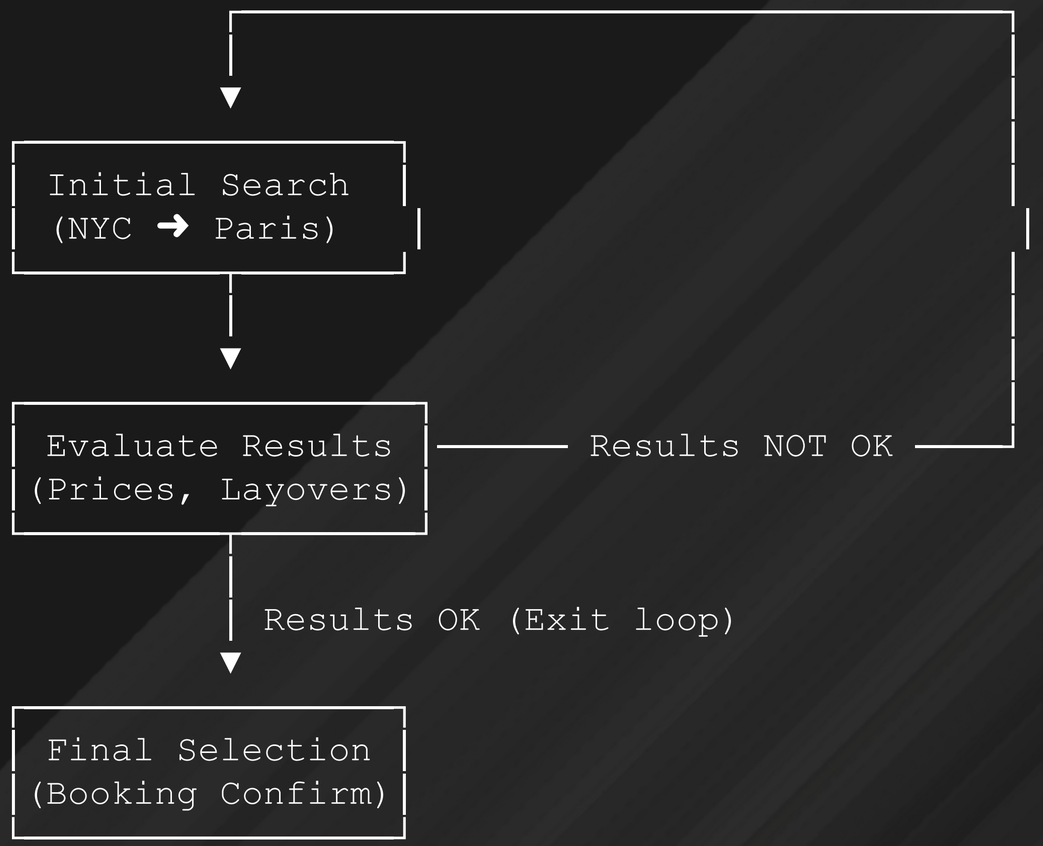
What is Agentic AI?

What is Agentic AI?

AKKA

- Agentic AI operates autonomously with minimal oversight.
- Proactively plans, executes, and adapts tasks.
- Dynamically responds to real-time changes.
- Autonomous, adaptable, and proactive.
- Capable of independent decision-making and problem-solving.
- Enables tools like ChatGPT to streamline user workflows.

Booking Travel (Ugh) The Feedback Loop



Explanation of Steps in the Diagram:

Initial Search:
Conduct initial travel search (NYC ➜ Paris).

Evaluate Results:
Check price, convenience, and flight details.

If Results NOT OK:
Adjust parameters (dates, price range, airports) and return to the initial search.

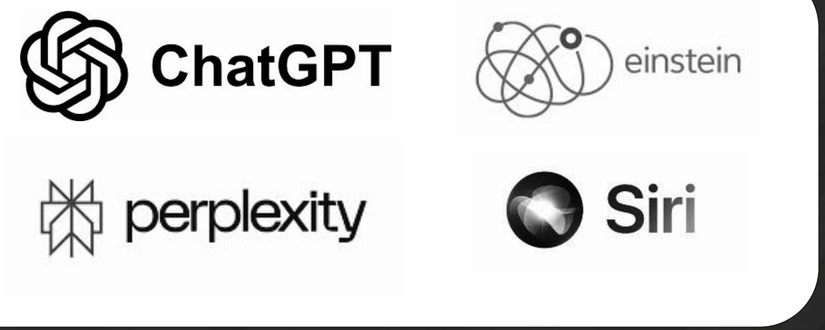
If Results OK:
Finalize selection, complete the booking, and exit the feedback loop.

AI is transforming our lives

AKKA

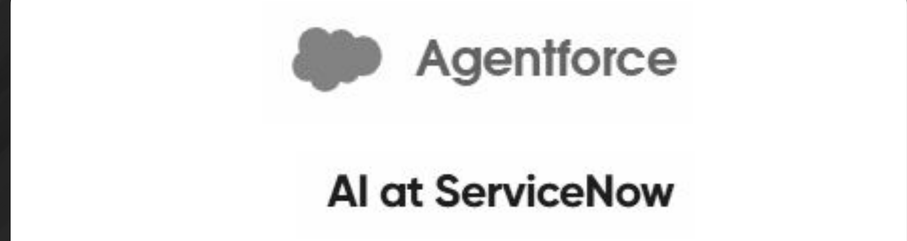
AI Assistant

A *user* app that understands natural language commands and uses a conversational AI interface to complete tasks on-demand.



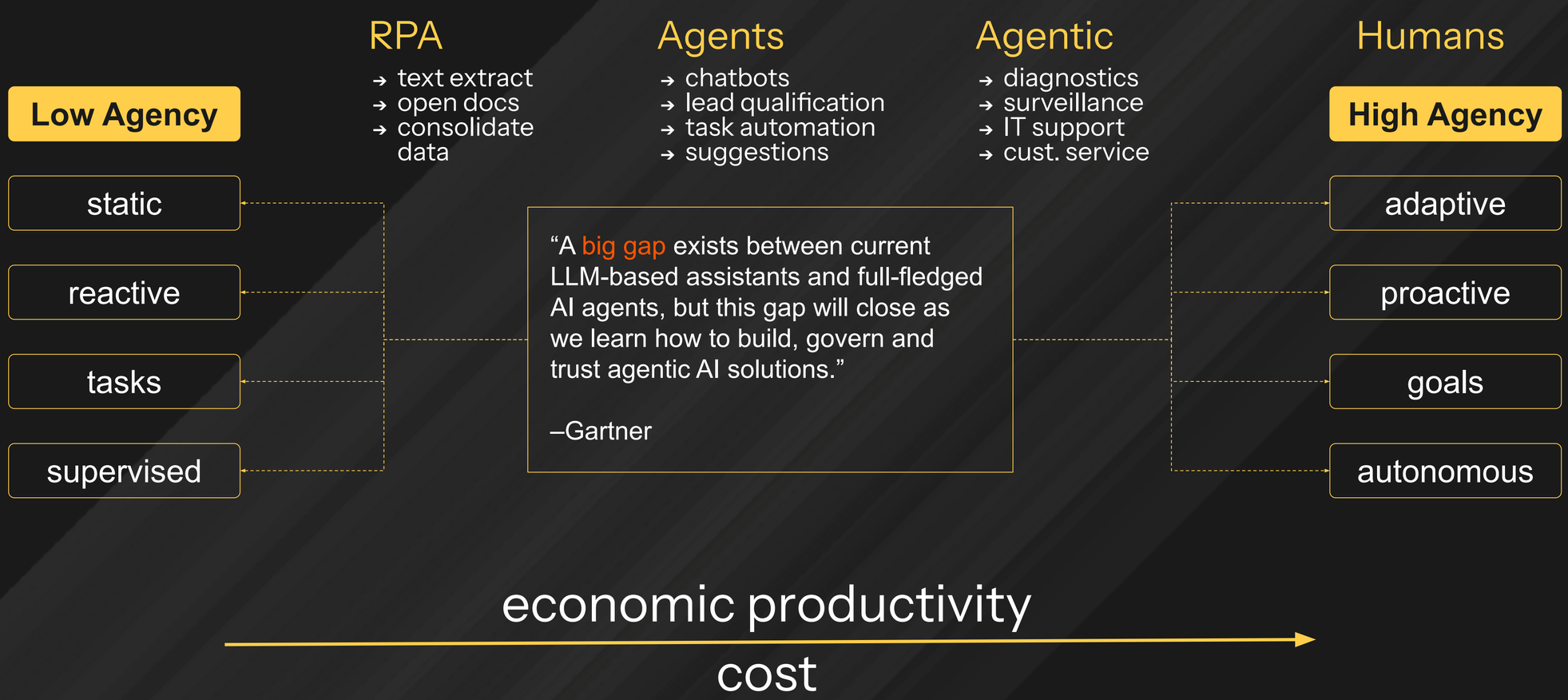
AI Agent

A *system* that can autonomously fulfill goals by interacting with other systems and agents.



AI Agency

Capacity to make meaning from your environment



AKKA

Agentic is the 5th wave of compute

Every human and device with dozens of sleepless assistants

	Mainframe	Web	Cloud	Mobile	Agentic
Users	thousands	millions	10 millions	billions	trillions
TPS	100	500	2,500	10,000	1,000,000
Order of magnitude growth		**5x**	**5x**	**4x**	**100x**

AKKA

A paradigm shift to AI-fueled app ecosystems

AI agents and apps become part of a symbiotic existence

By 2028, 33% of enterprise software applications will include agentic AI, up from less than 1% in 2024.

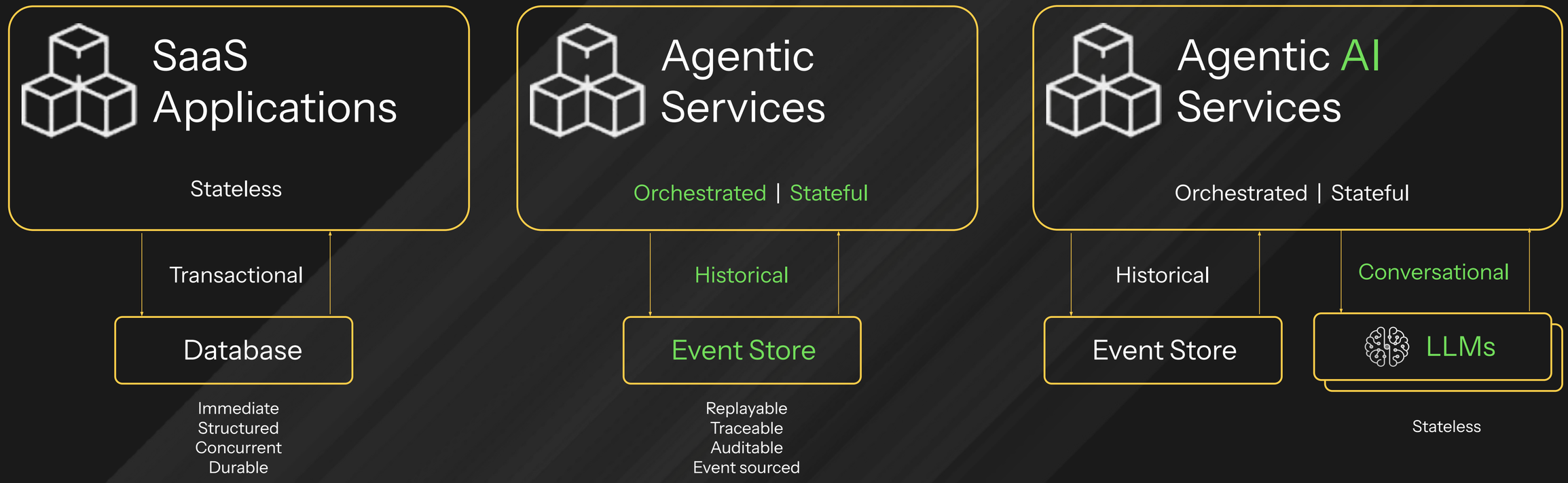
Gartner, *TSP 2025 Trends: Agentic AI — The Evolution of Experience, 24 February 2025*

The Challenges with Agentic AI

AKKA

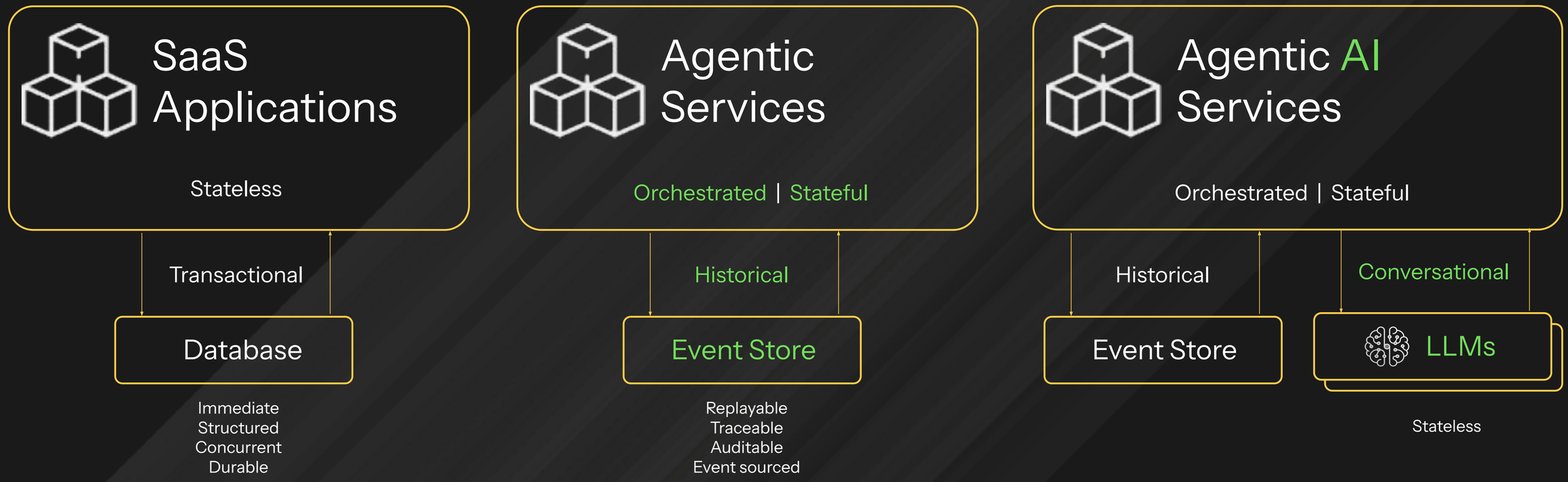
The Agentic Loop

A fundamental shift from request-response to contextual iterations



Transactional apps → Conversational agents

A fundamental shift from request-response to contextual iterations



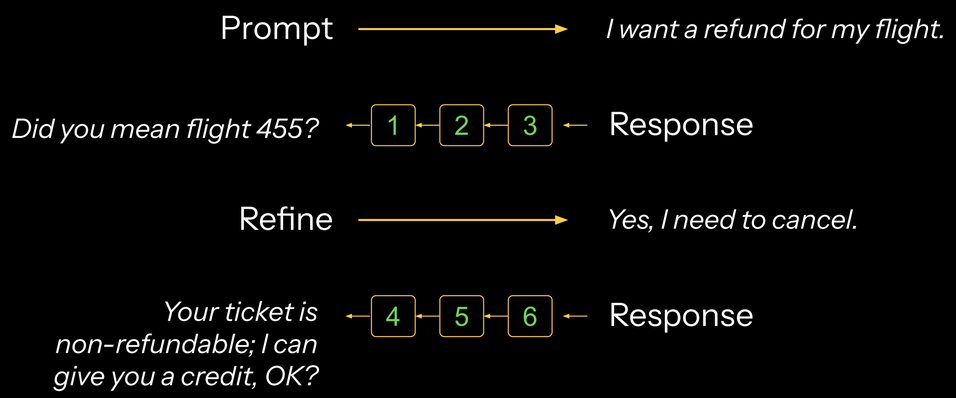
AKKA

Conversations are stateful

Context and conversation database now a part of the agentic stack

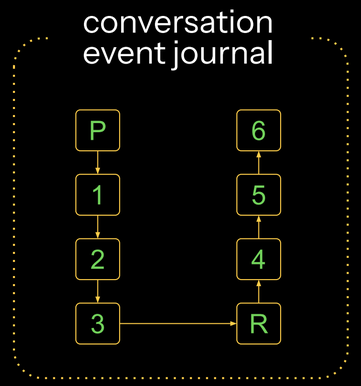
Conversations are ongoing

each iteration
adds context



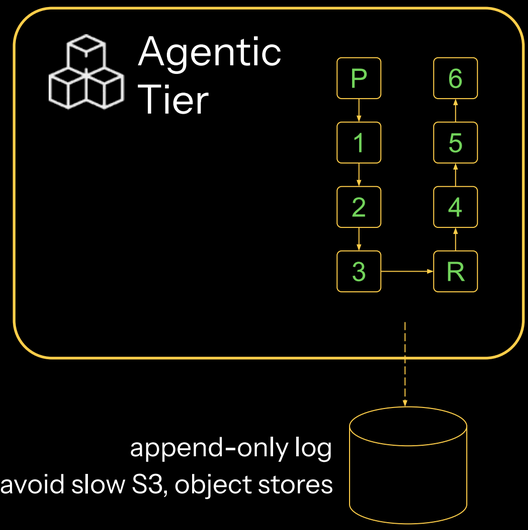
Conversational sessions

journaled sequences
for context and recovery



Conversational persistence

in-memory, durable journals
for speed + resilience



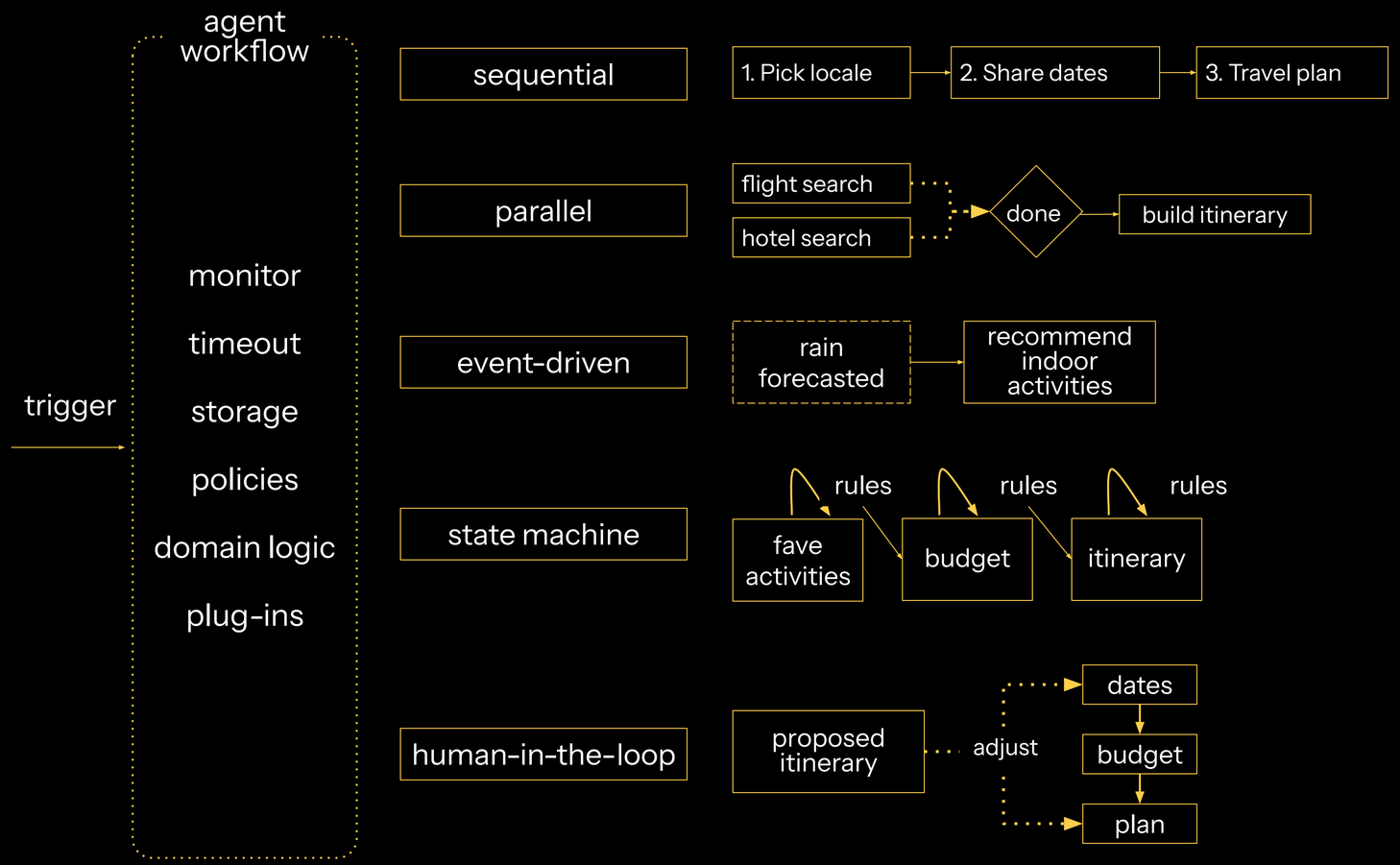
Agents are orchestrated services

AKKA

Workflows: traceable, auditable, debuggable, with point-in-time recovery

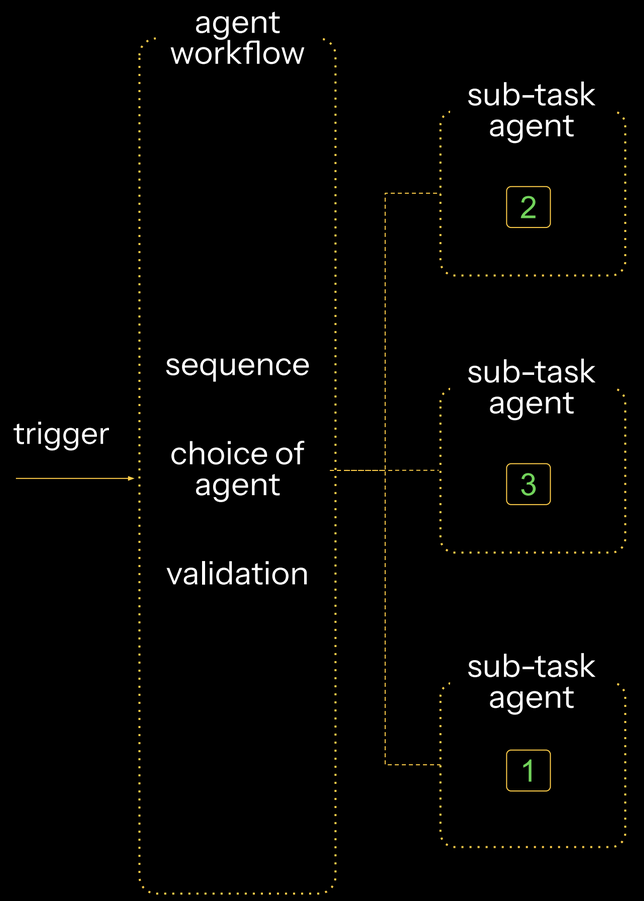
Agents are workflows

reliable execution of AI tasks with visibility into request / response data, built-in retries, and error compensation



Task chaining

AI agents break complex workflows into smaller, composable steps



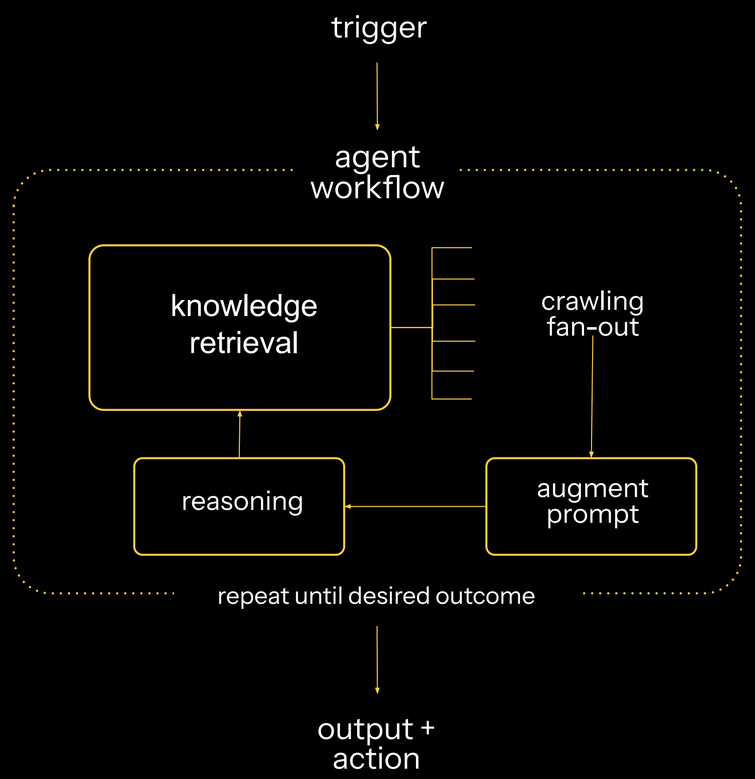
AKKA

Agent types orchestrate levels of agency

De-coupled, event-driven patterns and control loops

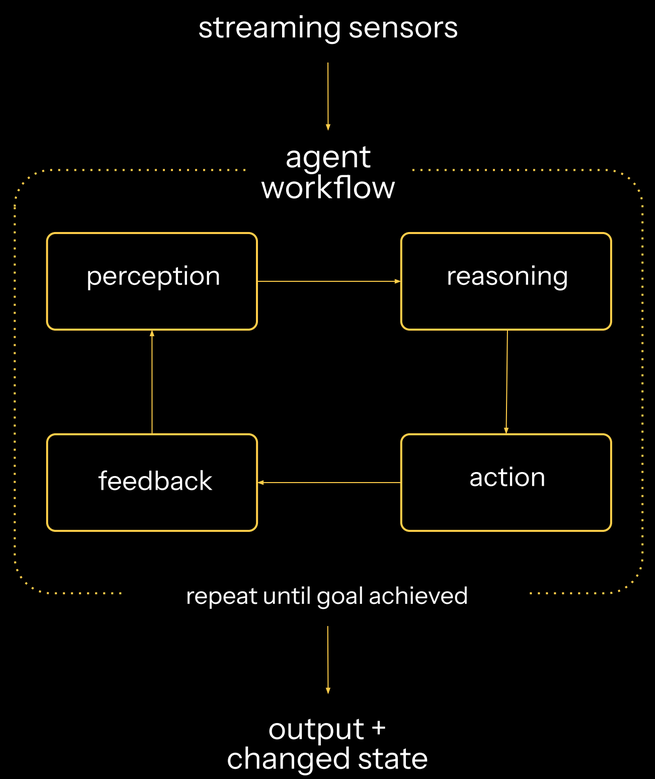
Retrieve - augment

agents that combine external knowledge with reasoning and action



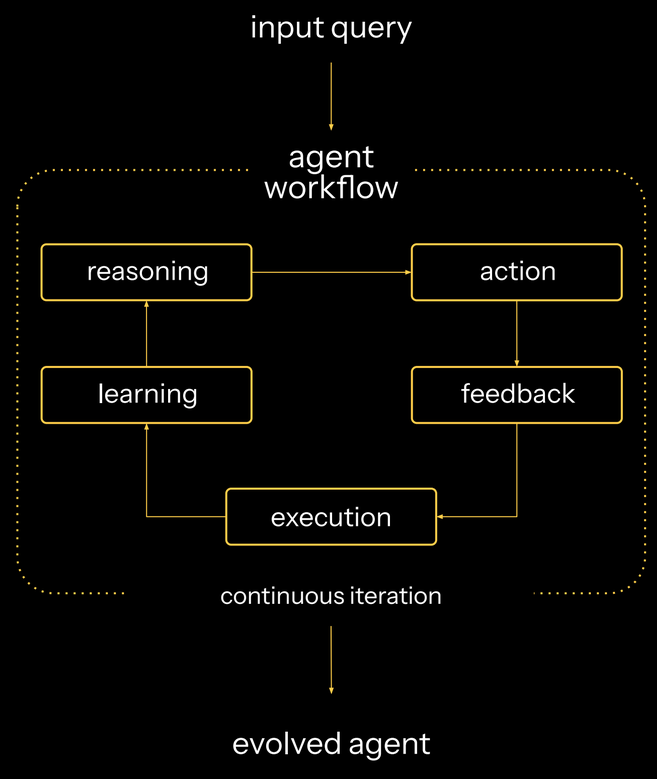
Environment controllers

control environments in real-time for robotics, edge, and automation



Self learning

agents that improve themselves over time through self-reflection and environment adaptation



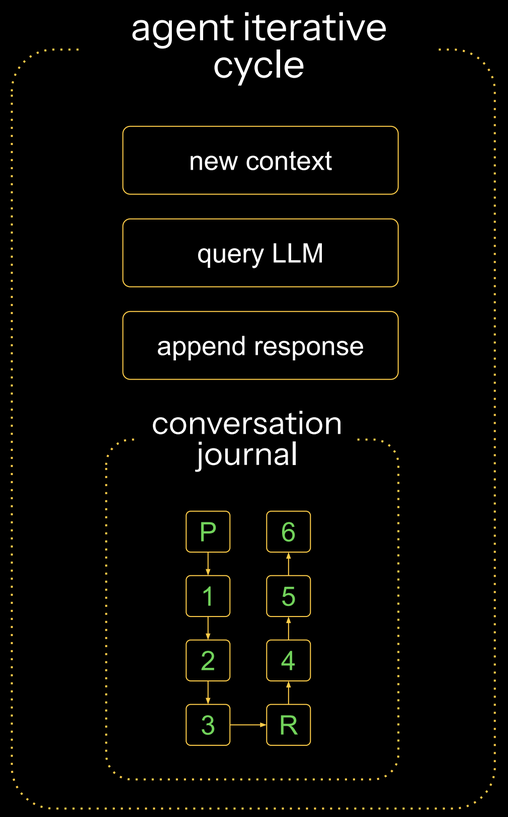
AKKA

Agentic AI augmentation cycle

Agents slow down on each iteration as context grows

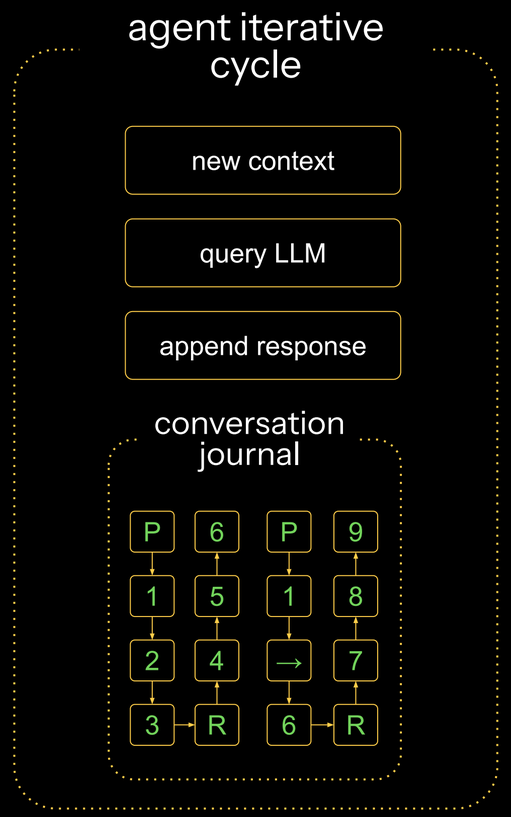
Agents start fast

small prompts, small conversations generate quicker responses



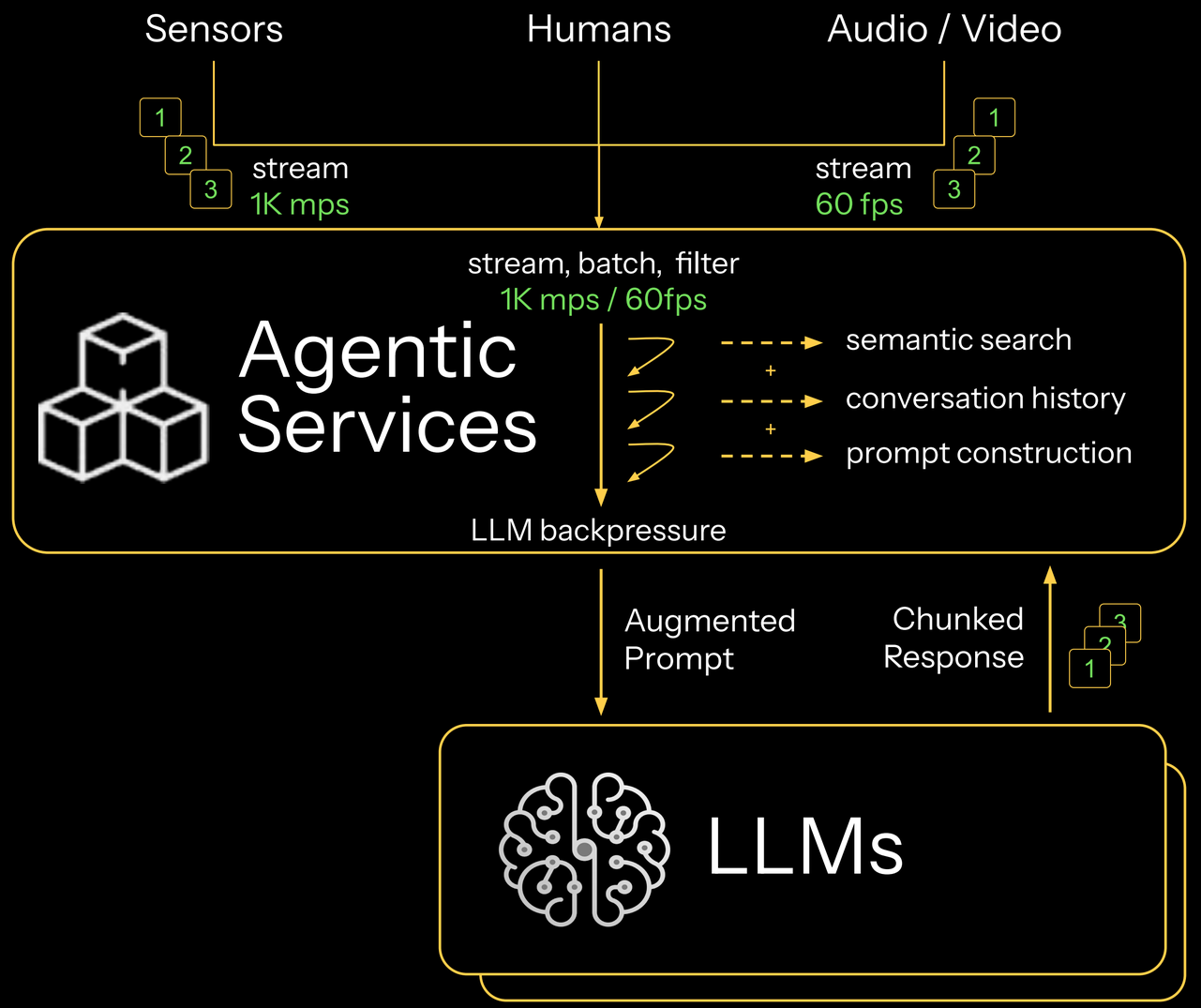
Agent iterations grow slower

Conversations and prompts grow, eventually hitting LLM token cap



Augment at streaming speeds

Agents augment from a continuous stream of inputs without overloading themselves or their LLMs



Agentic scale requires efficiency

More txs: each slower, less predictable and more costly

	SaaS	Agentic
Users	billions	20x
TPS	10,000	100x
p(99) Latency	10-80ms	15-400x
Cost / LLM tx	cheap	10-10,000x

Mar 25: the best performing LLM @ 86% MMLU accuracy costs $98 / 1M tokens, or ~850,000x more expensive than the average database transaction. The worst performing LLM @ 36% MMLU accuracy costs $.01 / 1M tokens, or 7x more expensive.

Bumpy path from POC to production

52%

fail to reach production

8+ months

POC to production

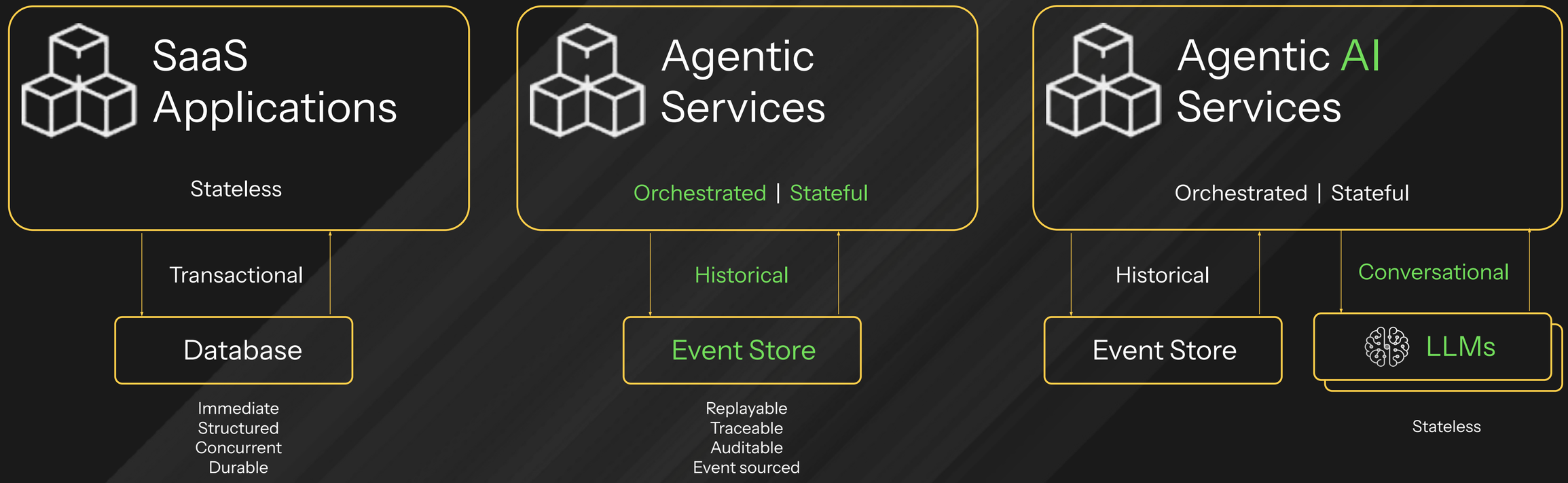
“Leaders reported that only 48% of AI POCs (Proof Of Concept) make it into production, and they take an average of 8.2 months to go from POC to production.”

Gartner

AKKA

Will Agentic AI Replace SaaS?

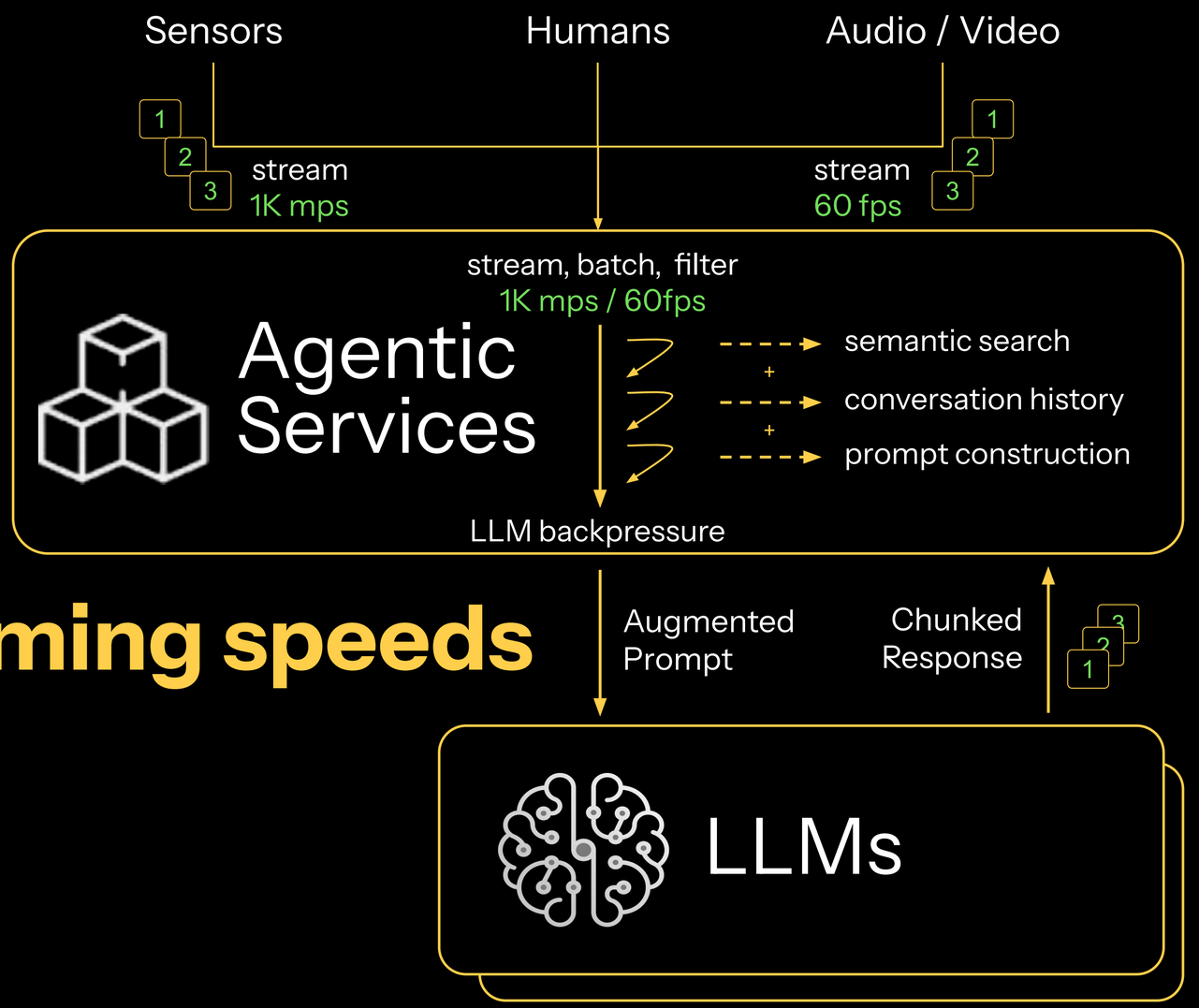
A fundamental shift from request-response to contextual iterations



Why not just chain SLM's together??

AKKA

Agents augment from a continuous stream of inputs without overloading themselves or their LLMs

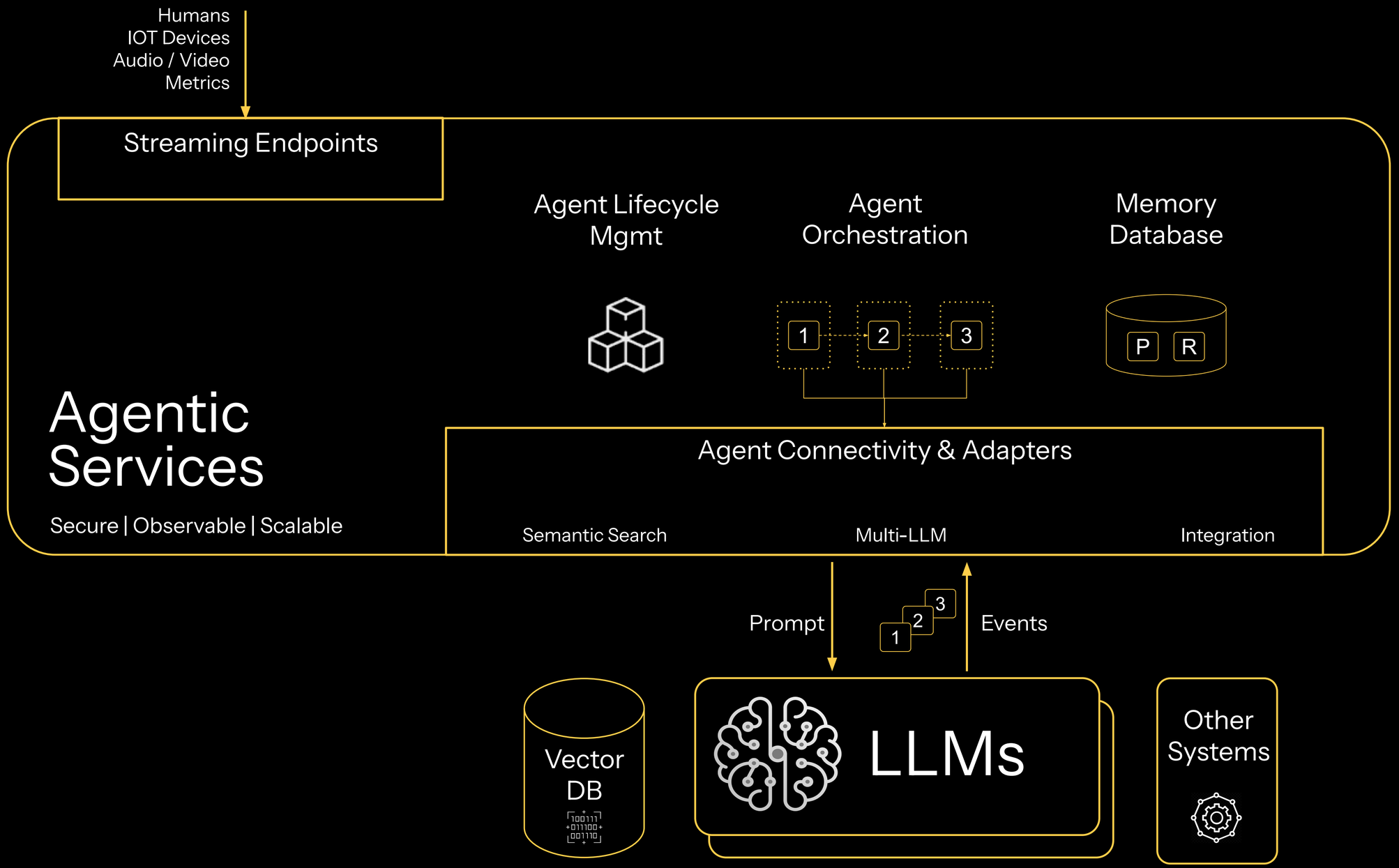


Augment at streaming speeds

AKKA

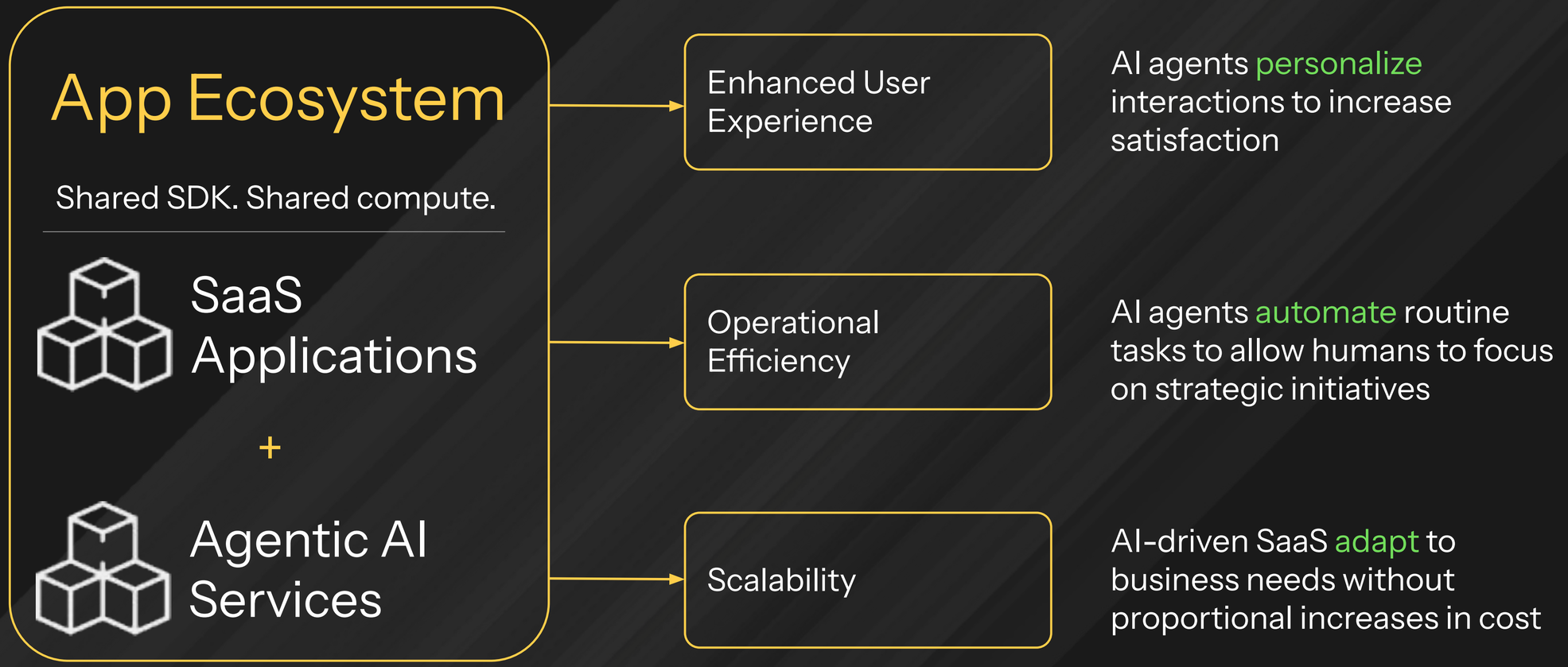
Blueprint for Agentic Services

AKKA


Humans
IOT Devices
Audio / Video
Metrics
Streaming Endpoints
Agent Lifecycle Mgmt
Agent Orchestration
Memory Database
1
2
3
P
R
Agentic Services
Secure | Observable | Scalable
Agent Connectivity & Adapters
Semantic Search
Multi-LLM
Integration
Prompt
Events
Vector DB
LLMs
Other Systems

AKKA

Agentic AI applications combine business logic, contextual history, and large language models (LLM) to continuously observe, reason, and act on behalf of users throughout workflows. They don't just respond to commands — they personalize interactions, automate routine tasks, and adapt to business needs.



The Akka agentic advantage

AKKA

- ✓ Agentic, AI, apps & data
- ✓ Hardened runtime
- ✓ Simple, expressive SDK
- ✓ Multi-region
- ✓ Automated ops

Streaming endpoints

- ➔ Shared compute: agentic co-execution with API services
- ➔ HTTP and gRPC custom API endpoints
- ➔ Custom protocols, media types, and edge deployments
- ➔ Real-time streaming ingest, benchmarked to over 1TB

Memory database

- ➔ Agentic sessions with infinite context
- ➔ Context snapshot pruning to avoid LLM token caps
- ➔ In-memory context sharding, load balancing, and traffic routing
- ➔ Multi-region context replication
- ➔ Replication filters for region-pinning user context data
- ➔ Embedded context persistence with Postgres event store

Agent connectivity & adapters

- ➔ Non-blocking, streaming LLM inference adapters with back pressure
- ➔ Multi-LLM selection
- ➔ LLM adapters & 100s of ML algos
- ➔ Agent-to-agent brokerless messaging
- ➔ 100s of 3rd party integrations

Agent orchestration

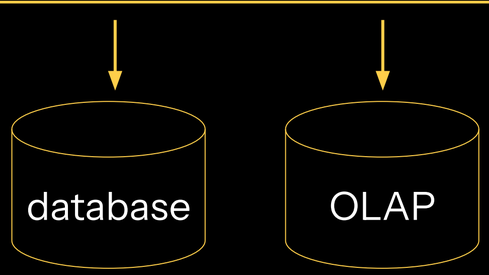
- ➔ Event-driven runtime benchmarked to 10M TPS
- ➔ SDK with AI workflow component
- ➔ Serial, parallel, state machine, & human-in-the-loop flows
- ➔ Sub-tasking agents and multi-agent coordination

Agent lifecycle management

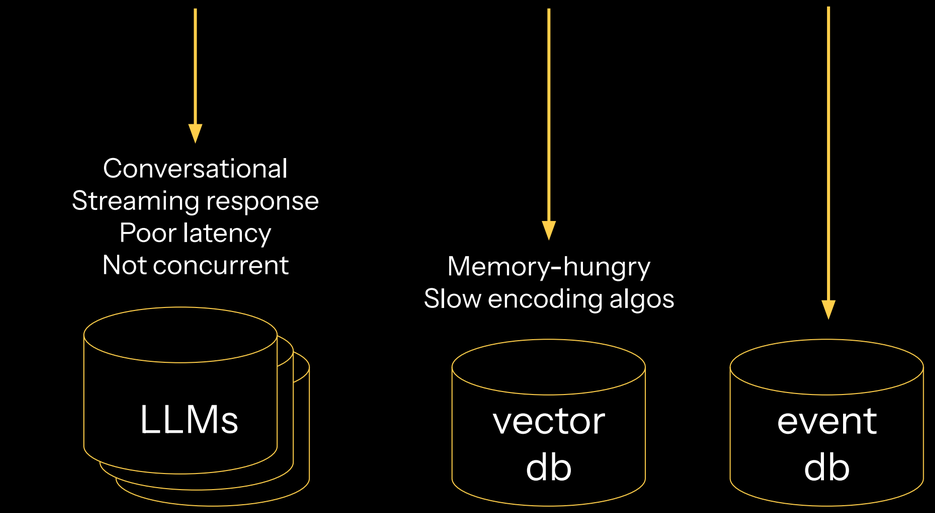
- ➔ Agent versioning
- ➔ Agent replay
- ➔ Event, workflow, and agent debugger
- ➔ No downtime agent upgrades

From n-tier to a-tier architecture

Humans + devices augmented with dozens of agent assistants that never sleep



TPS	thousands	hundreds
p(99) latency	10-50ms	50-300 ms



TPS	100x	5x	100x
p(99) latency	150 - 3000ms	50-200ms	5-150ms

Agentic is real
Let’s make it real for you

concept

proof

48 hours