Unlocking 10x larger AI models on existing hardware

Democratizing access to trillion-parameter AI through intelligent storage tiering.

Transforming commodity hardware into AI powerhouses.
Train 200B parameter models on a $1500 workstation

Awase Khirni Syed Ph.D.

Cache Fusion AI

Democratizing large-scale AI through Intelligent Memory Expansion

DRAFT VERSION

The Problem

$50B AI Memory Crisis

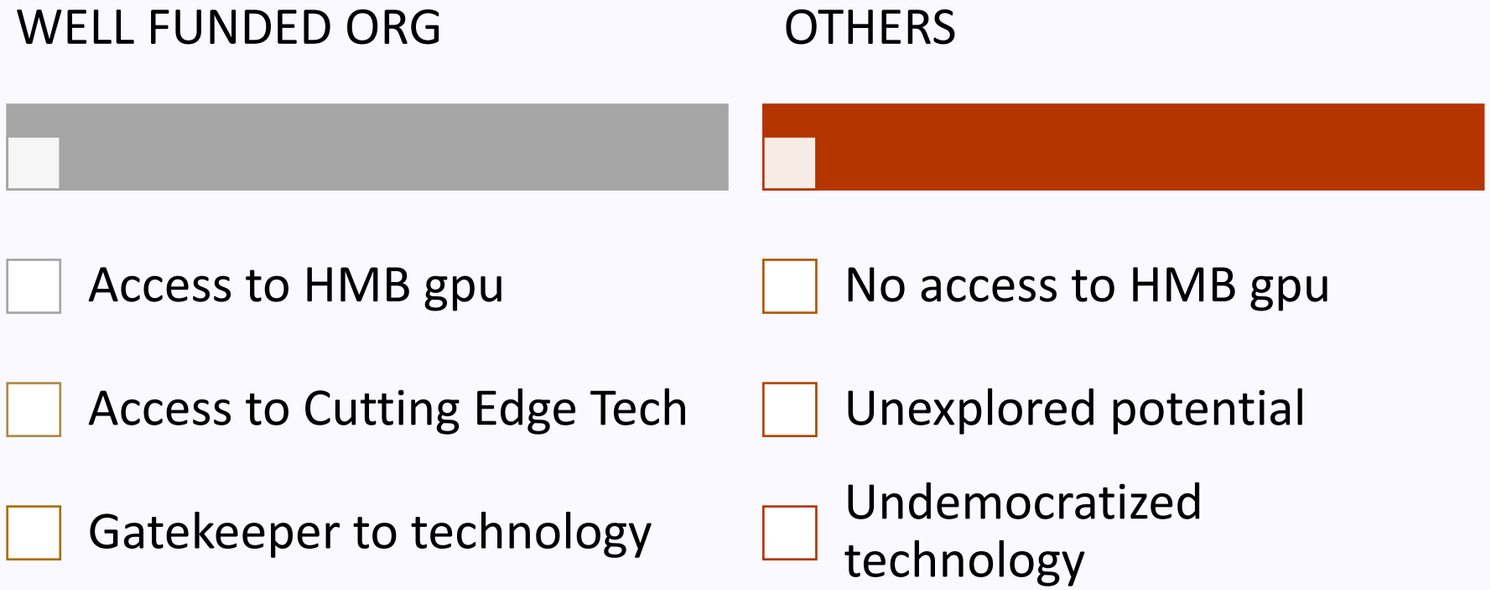
The gap between model growth and hardware capacity

Today’s AI landscape is defined by a critical constraint: Memory Limitations

- Prevent breakthrough research at accessible price points.
- Training large language models require hundreds of thousands of dollars in specialized hardware
- Creates an innovation barrier that excludes most researchers, startups, and institutions.

The result?

- A two-tier AI ecosystem where only well-funded organization can push boundaries, while brilliant ideas remain unexplored due to cost barriers



The Problem

$50B AI Memory Crisis

The gap between model growth and hardware capacity

AI progress is literally memory-bound. The gap between model growth and hardware capacity is creating a critical bottleneck that’s excluding 99.9% developers from the state-of-the-art AI.

10X

Model Growth

AI models growing annually—from 10M to 100B to 1T+ parameters

1.5X

GPU Memory

Annual hardware expansion—creating an exponentially widening gap

$50B+

Market Opportunity

Total addressable market for memory expansion solutions alone

Training 1T Parameter Model

Cost: $3M+ (H100 Cluster)

Only accessible to top 0.1% of companies

Fine-tuning 70B Model

Cost: $500K+

Limited to top 1% of companies

Research-scale AI

Cost: $100K+

University labs struggle to compete

Startups struggle to compete

The AI revolution is stalling not from lack of ideas, but from lack of accessible memory. Meanwhile, a $30,000 GPU offers 80 GB while a $300 NVMe SSD provides 4TB of storage.

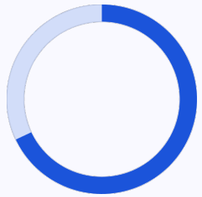
The Approach

CacheFusion

Intelligent Memory Expansion

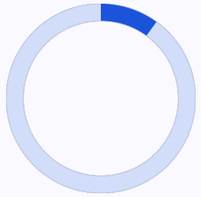
CacheFusion transforms affordable NVMe SSDs into expanded AI memory through filesystem-aware parameter management with ML-optimized caching algorithms.

Built on bcachefs with AI-specific optimization, we deliver 10-100x effective memory capacity at a fraction of traditional GPU costs. Open-Source License similar to DeepSeek.



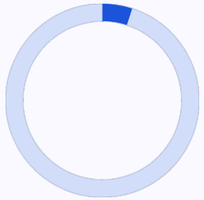
Native Performance

Of native VRAM speed maintained with intelligent caching



Memory Expansion

Effective capacity increase over traditional GPU configurations



Access Latency

For cached parameters with predictive prefetching

Transparent integration with PyTorch, TensorFlow, and JAX means no code changes required. Deploy/Install CacheFusion and immediately unlock larger model capacity on your existing hardware. Require High-end Gaming System Configuration with minimum 8GB gpu and 128 GB System RAM.

The Problem

$50B AI Memory Crisis

Targeting Four High-Value Customer Segments

Academic & Research (40%)

University AI/ML labs, graduate researchers, government facilities, and independent scientists with limited budgets seeking breakthrough capabilities.

Startups & SMEs (30%)

Pre-Series A AI/ML startups, mid-market companies adopting AI, consulting firms, and digital agencies requiring cost-effective scaling.

Enterprise R&D (20%)

Fortune 500 innovation labs, financial quant teams, healthcare research divisions, and manufacturing AI groups exploring new frontiers.

Cloud Providers (10%)

GPU cloud providers, AI-as-a-Service platforms, system integrators, and managed service providers seeking differentiated offerings.

Opportunity

Value Proposition

10X Larger Models at 1/10th the Cost

Cost Democratization

Enable $1,500 workstations to train **200B parameter models**. Reduce AI research entry cost from $100K+ to under $2,000.

Performance Accessibility

Achieve **13% of H100 performance** at just 0.6% of the cost. Scale linearly with additional NVMe drives.

Seamless Integration

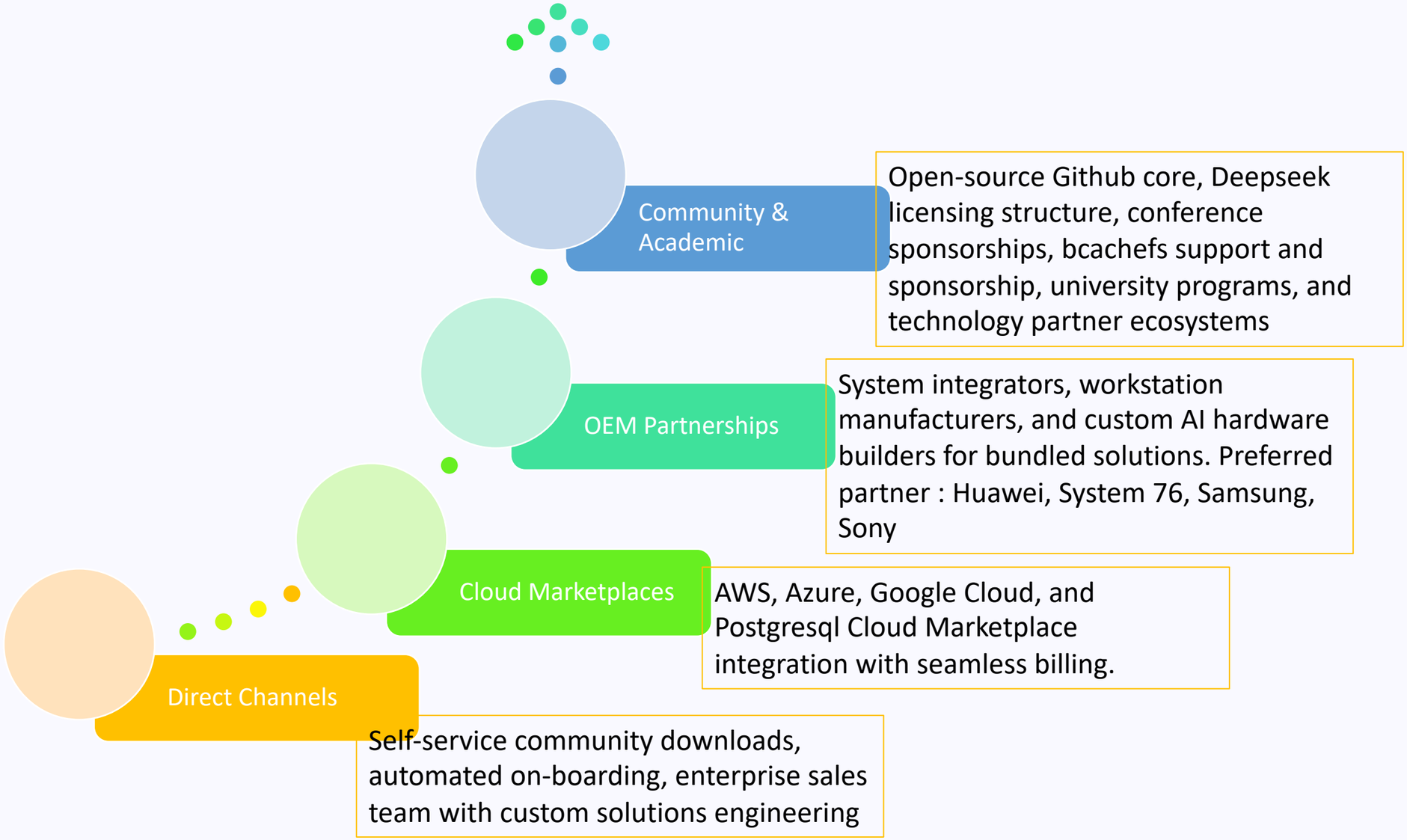
Framework-native with **PyTorch, TensorFlow, JAX**. Zero code changes required with automatic tier management.

Future-Proof Architecture

Hardware-agnostic design improves with storage technology. Open-core with enterprise support options.

Opportunity

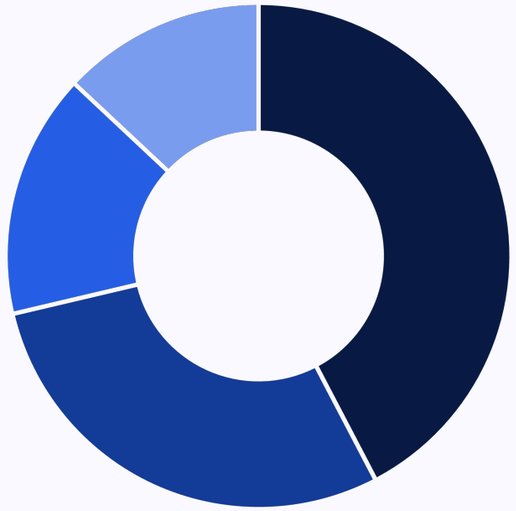
Multi-Channel Go-to-Market Strategy



Opportunity

MARKET

Intelligent Memory Expansion



Why Now?

- **Timing:** AI memory crisis at inflection point
- **Technology:** bcachefs merged into Linux kernel 6.7+
- **Market:** NVMe prices at all-time low ($50/TB)
- **Demand:** Open-source LLM explosion creating urgent need

01

Research & Academia

5,000+ university AI labs globally, government research institutions, independent researchers seeking accessible AI infrastructure

02

Startups & SMEs

20,000+ AI/ML startups, mid-market companies adopting AI, consulting firms delivering AI services

03

Enterprise AI

Fortune 500 AI initiatives, financial services AI deployment, healthcare and biotech research applications

04

Cloud & Service Providers

AI-as-a-Service platforms, GPU cloud providers, MLops platforms expanding capabilities

Opportunity

Business Model

Intelligent Memory Expansion

Community Edition

10$ per annum

1 Terabyte memory expansion

Community support

Open-source integrations

Target: Researcher & Students, General Public

Pro Edition

999$ per annum

2 Terabyte memory expansion

Community support

Open-source integrations

Performance analytics

Priority Support

Advanced Optimizations

Target: Researcher & Students, General Public, AI Startups, Research Labs

Enterprise

6000 $/year for 50 licenses

10000 $/year for 100 licenses

15000 $/year for 200 licenses

SLA guarantees

Custom integrations

Training & Certification

Target:Fortune 500

OEM Licensing

% of hardware

Pre-installed solutions

Customized implementations

Co-branding options

Target: System Manufacturers (Huawei, Asus)

Value-Based Pricing: Enterprise tier priced at 30% of infrastructure cost savings, ensuring immediate ROI for customers while capturing value proportional to impact delivered.

A fraction of our earning goes to support bcachefs project and tools that support it.

Opportunity

Strategic Partnerships Driving Growth

Intelligent Memory Expansion

Technology Partners

NVIDIA Strategic Alliance

CUDA/cuDNN optimization, developer program participation, joint marketing for complementary positioning.

AI Framework Teams

Deep integration with PyTorch Foundation, TensorFlow (Google), and JAX teams for native support.

Linux Distributions

Canonical (Ubuntu), Red Hat/Fedora, SUSE partnerships for seamless OS integration.

Hardware & Channel Partners

OEM Manufacturers

Huawei, Samsung, Sony, System 76, and custom AI hardware builders for bundled offerings.

Cloud Providers

AWS, Azure, Google Cloud instance optimization and marketplace integration.

Research Institutions

Category

Competitive Advantage

Intelligent Memory Expansion

Solution	Expansion	Performance	Cost	Integration	Winner
CacheFusion	10-100X	60-80%	💰	⭐⭐⭐⭐⭐	✓
NVIDIA H100	1X	100%	💰💰💰💰💰	⭐⭐⭐⭐⭐	
CPU Offloading	5-10X	10-30%	💰💰	⭐⭐	
Model Parallelism	N GPUs	60-80%	💰💰💰💰💰	⭐	
Quantization	2-4X	85-95%	💰	⭐⭐⭐	

No Hardware Lock-in

Works with any GPU plus NVMe configuration—vendor agnostic and future-proof

3+ Year Technical Lead

Deep bcachefs development expertise with AI-specific patents pending

90% Cost Savings

Deliver trillion-parameter capabilities at fraction of traditional infrastructure costs

Category

Competitive Advantage

Defensible Technology & Intellectual Property

Core Competitive Advantages

Patent Portfolio Protection

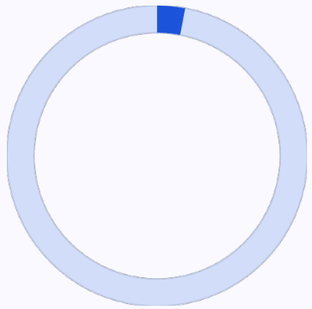
3 filed, 2 pending patents covering AI-optimized caching algorithms, filesystem memory expansion, and distributed training optimizations.

Open-Source Strategic Moat

GPL-licensed bcachefs enhancements with MIT/Apache framework integrations create community adoption while protecting commercial extensions.

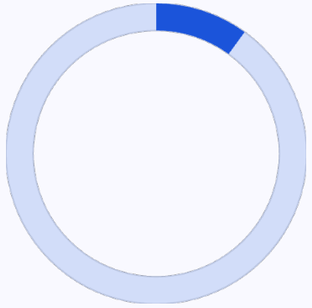
Academic Validation

Stanford and MIT endorsements, conference presentations at NeurIPS and ICML, plus production deployment testimonials.



Filed Patents

Core technology protection



Core Engineers

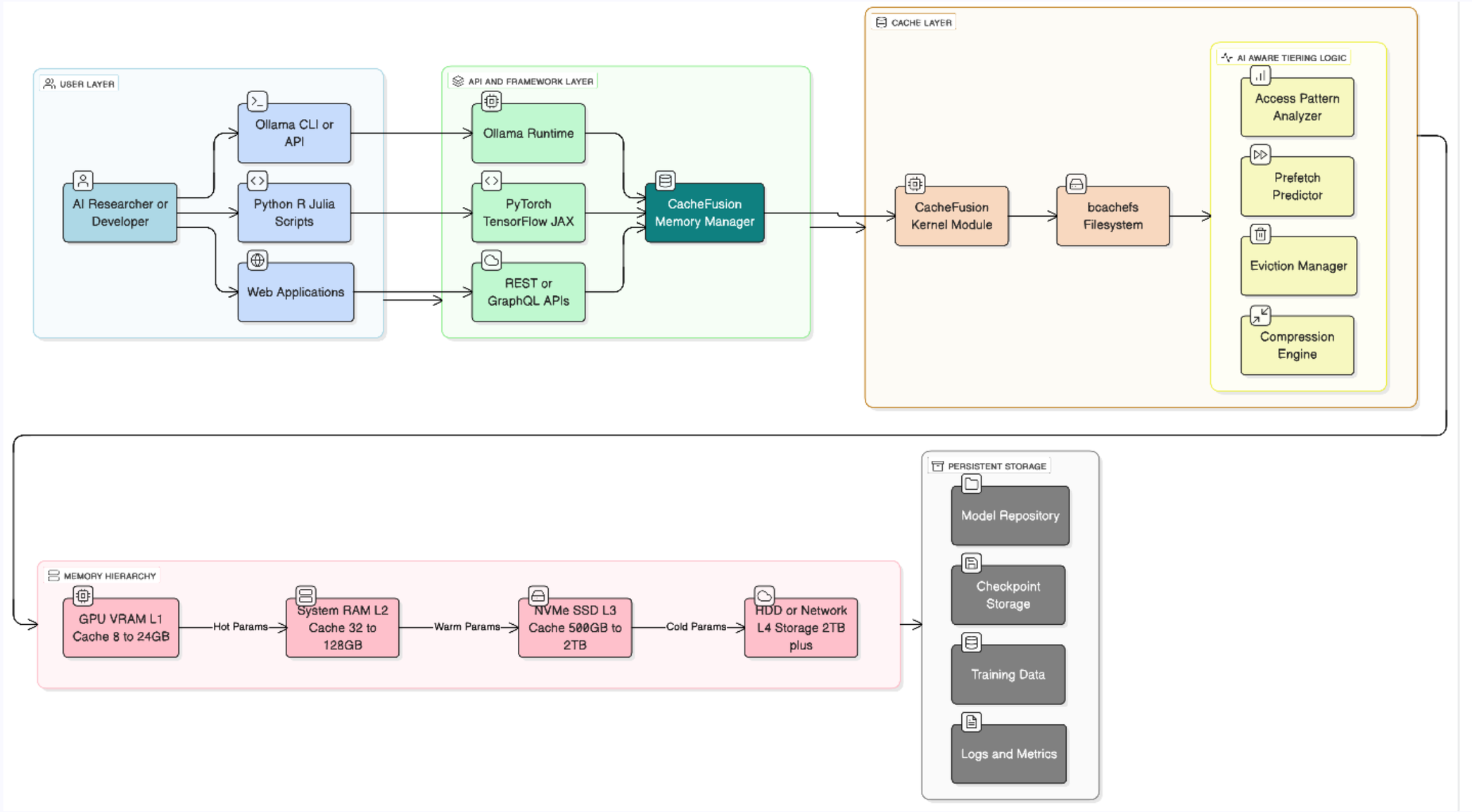
Linux kernel + AI experts

Software Margins

Highly scalable licensing

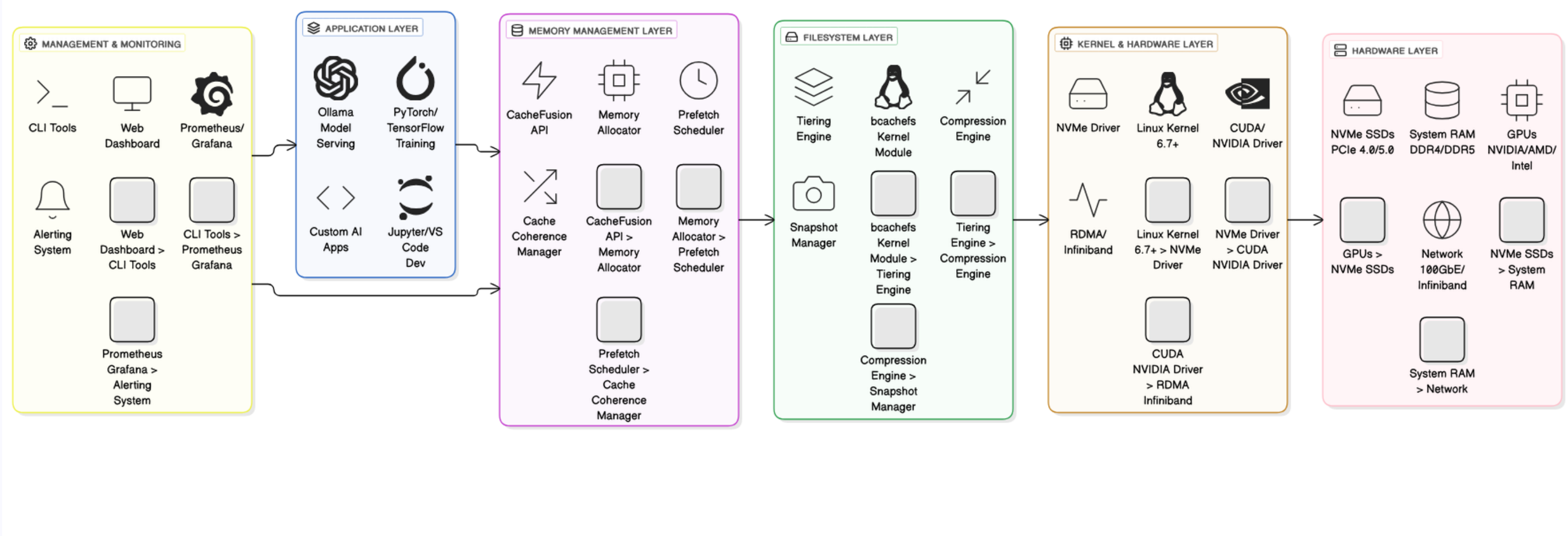
Category

Conceptual Architecture



Category

Software Architecture Stack



Category

Go-To-MARKET

Go-to-Market: Three-Phase Expansion

Phase 1: Beachhead

Months 1-6

Target AI research labs and PhD students with free Community Edition. University partnerships drive adoption.

Goal: 1,000 active deployments

Phase 2: Scale

Months 7-18

Focus on AI startups and mid-market companies via cloud marketplace with performance-based pricing.

Goal: $2M ARR, 50 enterprise customers

Phase 3: Enterprise

Fortune 500 AI teams through direct sales and OEM partnerships. Industry-standard

Goal: $20M ARR, market leadership

1 2 3

Channel Strategy

- Direct enterprise sales
- Cloud marketplaces
- OEM pre-installation
- Open-source community

Key Partnerships

- NVIDIA/CUDA ecosystem
- PyTorch/TensorFlow/JAX teams
- SSD manufacturers
- System integrators

Unit Economics

- CAC: $5K enterprise
- LTV: $50K enterprise
- 90% gross margin
- 6-month payback

ROADMAP

ROAD AHEAD

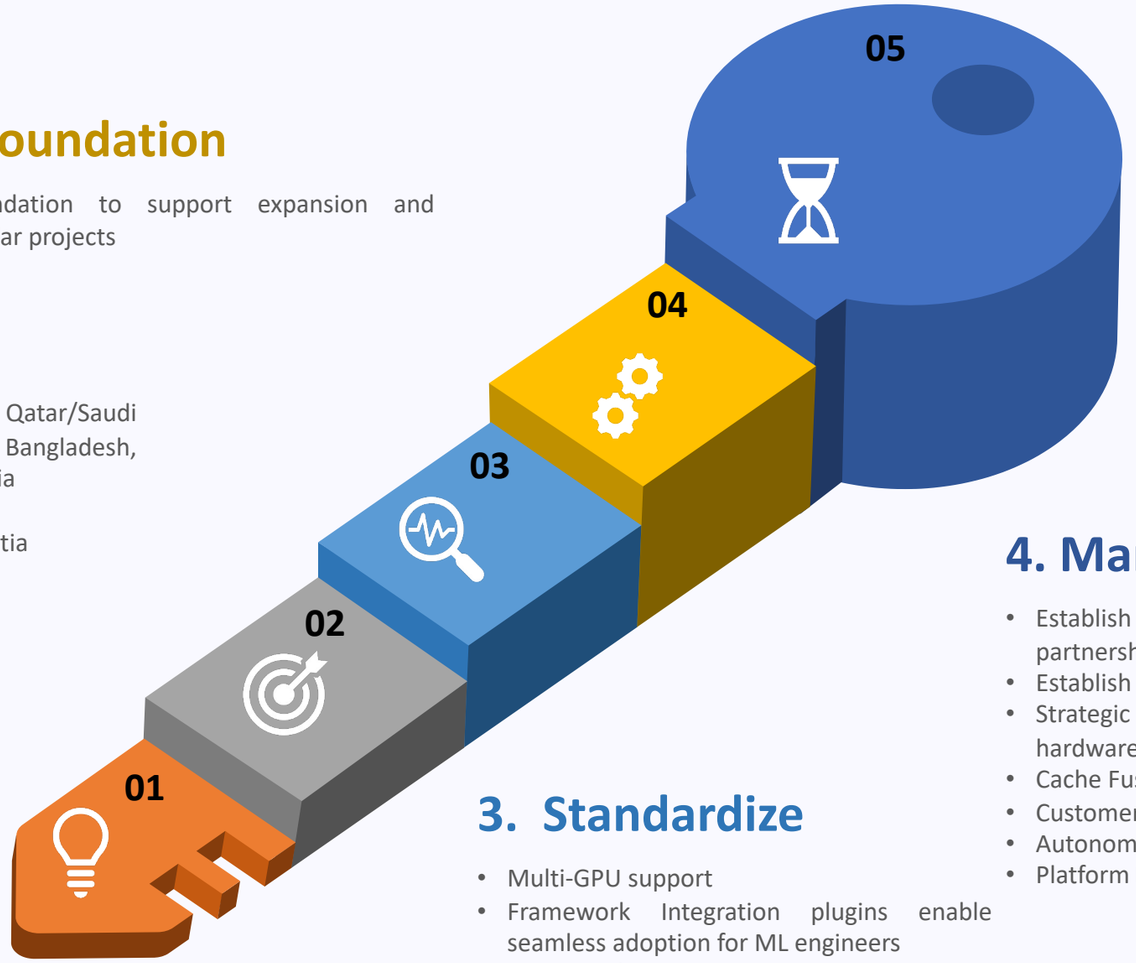
Democratizing AI

Vision: Build the definitive platform for AI memory expansion, enabling trillion-parameter models on commodity hardware

Mission: Democratize access to large-scale AI by making expensive GPU memory constraints obsolete

5. Foundation

- Foundation to support expansion and similar projects



2. Scale

- Expand to 10 geographical regions Brazil, Spain, Qatar/Saudi Arabia/Oman/Iran, China, Pakistan, India, Bangladesh, Malaysia, Indonesia, Japan, Newzealand/Australia
- Reac
- Create Challenges and Solutions with the consortia
- Achieve operational profitability.
- CacheFusion Core v0.1
- 1000+ github stars for each month

1.Launch & Prove

- Validate market need with 1000+potential customers in 5 customer segments
- Assemble core team of 25 Ph.D. in relevant streams
- Create consortium of university and startups interested to participate
- Raise $2M funding through crowdfunding
- Use Deepseek model license, open source
- Pledge 20% revenue to bcachefs for support
- Work with PyTorch, Jax and Tensorflow teams for integration with bcachefs
- POC to achieve 50% H100 performance at 5% cost using bcachefs + ollama

3. Standardize

- Multi-GPU support
- Framework Integration plugins enable seamless adoption for ML engineers
- Enterpise-beta 100 paying customers
- CacheFusion Cloud
- AI Caching Algoirthms
- Cross-Platform support for MacOS and Windows.
- Distributed Training

4. Market Adoption

- Establish global employment and health care partnerships
- Establish global educational partnerships
- Strategic Partnerships with cloud providers, hardware partners
- Cache Fusion 2.0
- Customer Kernel Module
- Autonomous Optimization
- Platform Ecosystem & Marketplace plugins

DREAM TEAM

TEAM

A powerhouse of technology researchers, technologists and software architects

[Awase Khirni Syed Ph.D.]

Chief Executive Officer & CTO

Strategic Technology Leader with 18+ years of experience driving digital transformation, solution architecture and enterprise-wide technology integration across financial, geospatial intelligence, geospatial search and technology sectors. Proven track record of leading cross-functional teams, implementing cutting-edge automation strategies, and delivering data-driven solutions that enhance operational efficiency and business outcomes. Expert in bridging technical solutions with business objectives and communicating complex concepts to diverse stakeholders. Proficient in managing and coaching/empowering large technical teams.