PURESTORAGE®

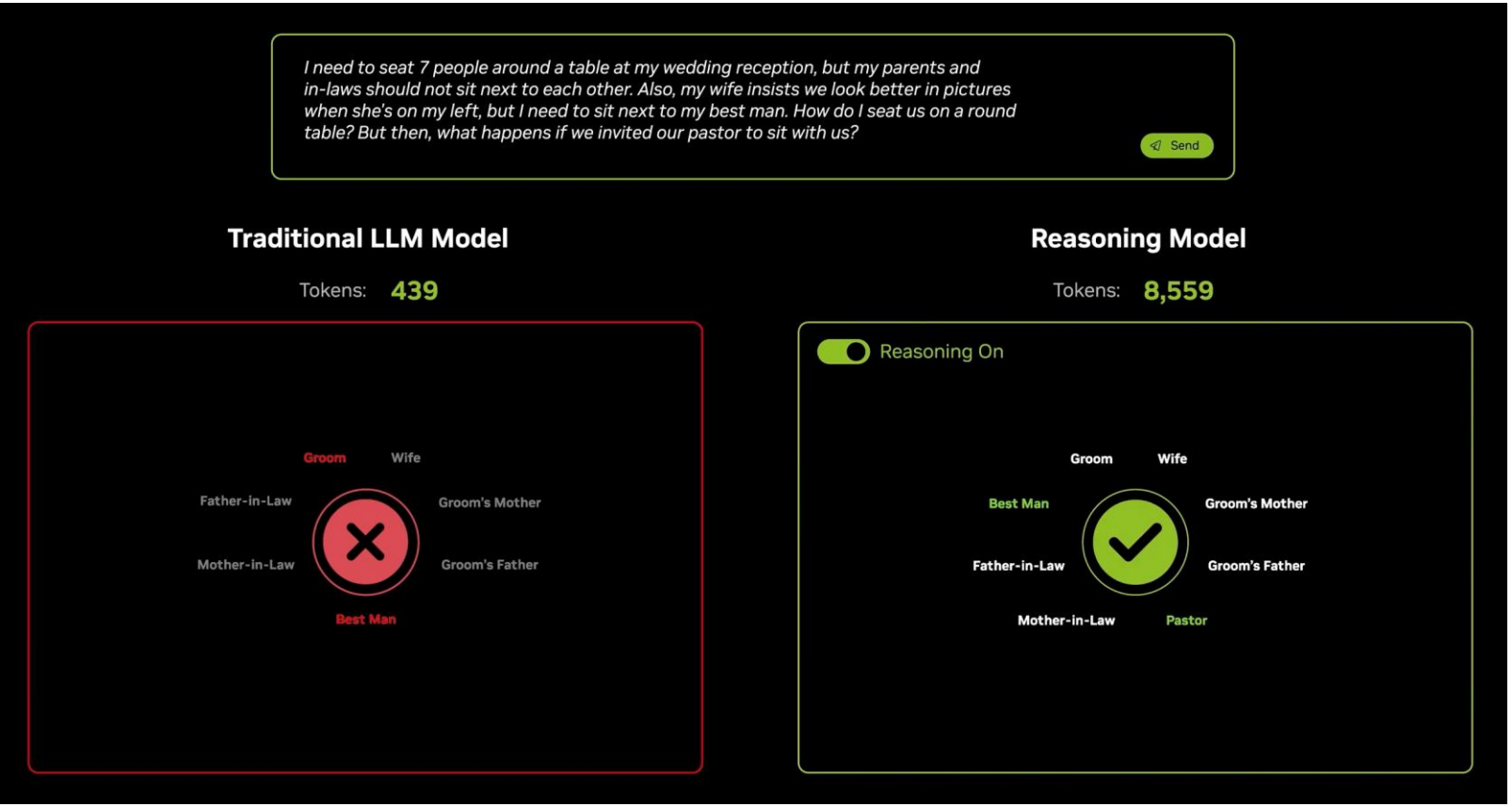
AI 혁신의 시작: 2025 GTC 인사이트와 함께하는 AI-Ready 인프라 및 데이터 통합 전략 해법

퓨어스토리지 코리아
강신우 이사
Consulting System Engineering

sean.kang@purestorage Apr 2025

GTC 2025 Keyword

Token, Reasoning, AI factory, Move to Ethernet, and others.



예측가능한 경우의 수는,
14,000,605 번의 패배, 그리고 1번의 승리

©2025 Pure Storage

GTC 2025 Keyword

Token, Reasoning, AI factory, Move to Ethernet, and others.

	Enterprise AI Workloads	AI Factories	Hyperscalers
Throughput	100GB/sec–1TB/sec	1TB/sec–50TB/sec	>50TB/sec
Capacity	50TBs–100PBs	100PB–Multiple EBs	10s to 100s of EBs
Total # of GPUs	<1,000 GPUs	1,000–10s of thousands of GPUs	> 10s of thousands of GPUs

- ***Infinity Scale.***
- ***AI-as-a-Service.***
- ***Reference Design.***

©2025 Pure Storage

GTC 2025 Keyword

Token, Reasoning, AI factory, Move to Ethernet, and others.

- ***Reasoning Model***
- ***Token***
- ***AI Cloud / AI Factory***
- ***Move to Ethernet***
- ***Enterprise Application → Enterprise AI***
- ***Energy Efficiency***

©2025 Pure Storage

GTC 2025 Keyword

Token, Reasoning, AI factory, Move to Ethernet, and others.

- ***Reasoning Model***
- ***Token***
- ***AI Cloud / AI Factory***
- ***Move to Ethernet***
- ***Enterprise Application → Enterprise AI***
- ***Energy Efficiency***

NVIDIA 가 이더넷으로 진출한 이유는?

이더넷을 Infiniband 성능과 동등하게 만들면(?)
기존 엔터프라이즈에서도 쉽게 사용하고 관리할 수 있기 때문.

→ 이더넷에 대한 높은 운영편의성과 비용성, 확장성 증명

시스코와의 협업

A Unified Architecture for AI Workloads

To date, most of the industry dialog on AI has been focused on chips, computing power, and Large Language Models. As enterprises begin to take on AI workloads, though, the challenge of connecting computing resources and data together – both inside of and in between data centers and clouds – is going to be the next frontier of AI innovation. In this sense, the network is going to be the key to successfully scaling out AI in the enterprise. Under the proposed partnership, here's how Cisco and NVIDIA intend to lead the charge:

1. NVIDIA will enable Cisco Silicon One coupled with NVIDIA SuperNICs to become part of the NVIDIA Spectrum-X Ethernet networking platform. Cisco would be the only partner silicon included in NVIDIA Spectrum-X. This is not only massive validation for Silicon One, but it also highlights the centrality of Cisco technologies – including silicon, optics, and software – in the AI technology landscape.

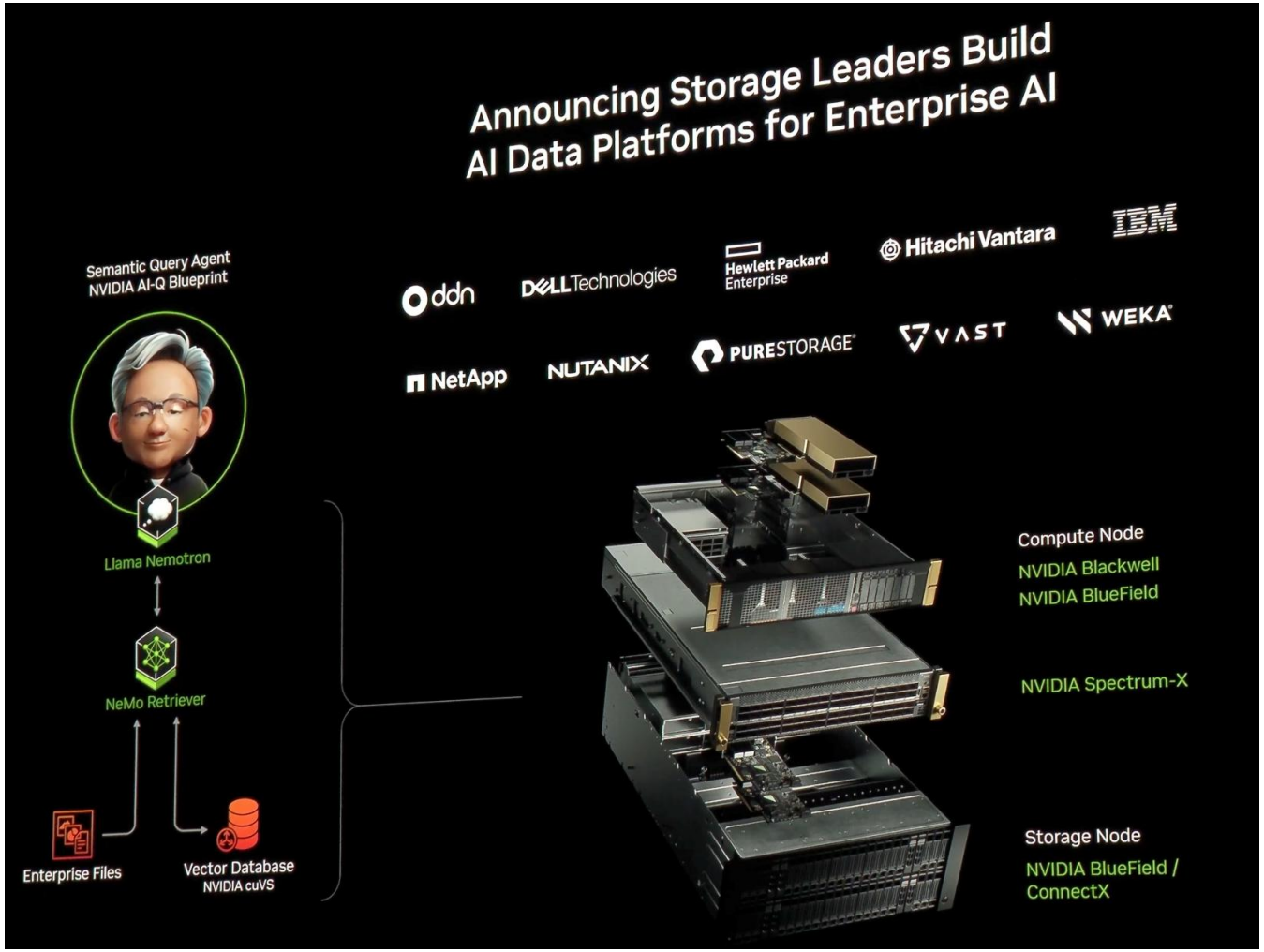
2. Cisco will be building systems that combine NVIDIA Spectrum silicon with Cisco OS software, allowing customers to simultaneously standardize on both Cisco networking and NVIDIA technology in the data center. Every enterprise will have a huge variety of use cases and applications for AI in their business. By enabling customers to leverage Nvidia Spectrum Silicon that's specialized for back-end connectivity or Cisco Silicon One, both under a common Nexus software stack (NX-OS and Nexus Dashboard), we will enable an exciting new level of interoperability in the industry.

https://blogs.cisco.com/news/cisco-and-nvidia-unite-to-propel-ai-into-the-enterprise

©2025 Pure Storage

GTC 2025 Keyword

Token, Reasoning, AI factory, Move to Ethernet, and others.



- *각 제조사 별 현재 진행중인 통합 프로젝트 소개*
- *NVIDIA Cloud Partner(NCP)*
- *FlashBlade//Exa*
- *Performance Scale*
- *S3 over RDMA(Fast Object)*

©2025 Pure Storage

AI 프로젝트 투자 고려사항

Training is just a small piece of a much larger & more complex picture

- **Clear Use Case**: *구체적이고 측정 가능한 사용사례*
- **Realistic Expectations**: *달성가능한 목표*
- **Data Readiness**: *양질의 데이터와 크기, 데이터 접근성 평가*
- **Infrastructure**: *AI 워크로드에 적합한 인프라인지 확인*
- **Executive Buy-In**: *기업의 C-Level 의 지원과 문화*

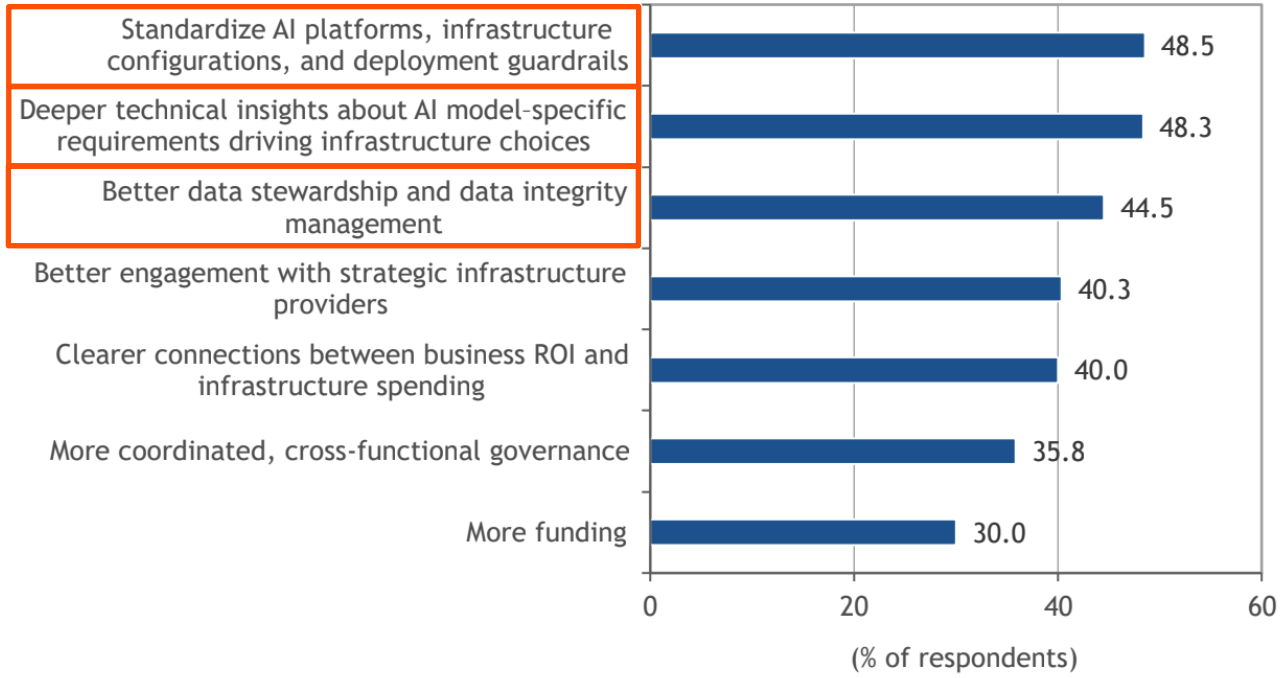
내부 고려사항 or 외부 고려사항

©2025 Pure Storage

AI-Ready Infrastructure

- 비용최적화
- 확장성
- 보안
- 효율성
- 상호 호환성

Q. *What are the most important things your organization needs to do to ensure its AI infrastructure is ready for business in 2026 and beyond?*



n = 600

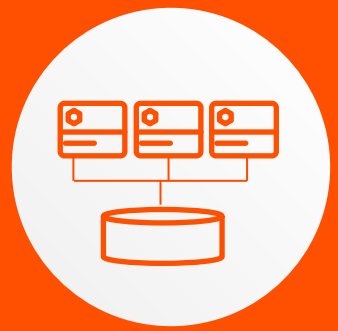
Source: IDC's *AI Infrastructure Survey,* September 2024

©2025 Pure Storage

AI-Ready Infrastructure 고려사항

AI Acceleration

AI 개발
속도 가속화



Dissagregated & Engineered

컴퓨팅, 스토리지
분리 구축 아키텍처

Deployment Consideration

클라우드, 온프렘,
VM, 컨테이너,,

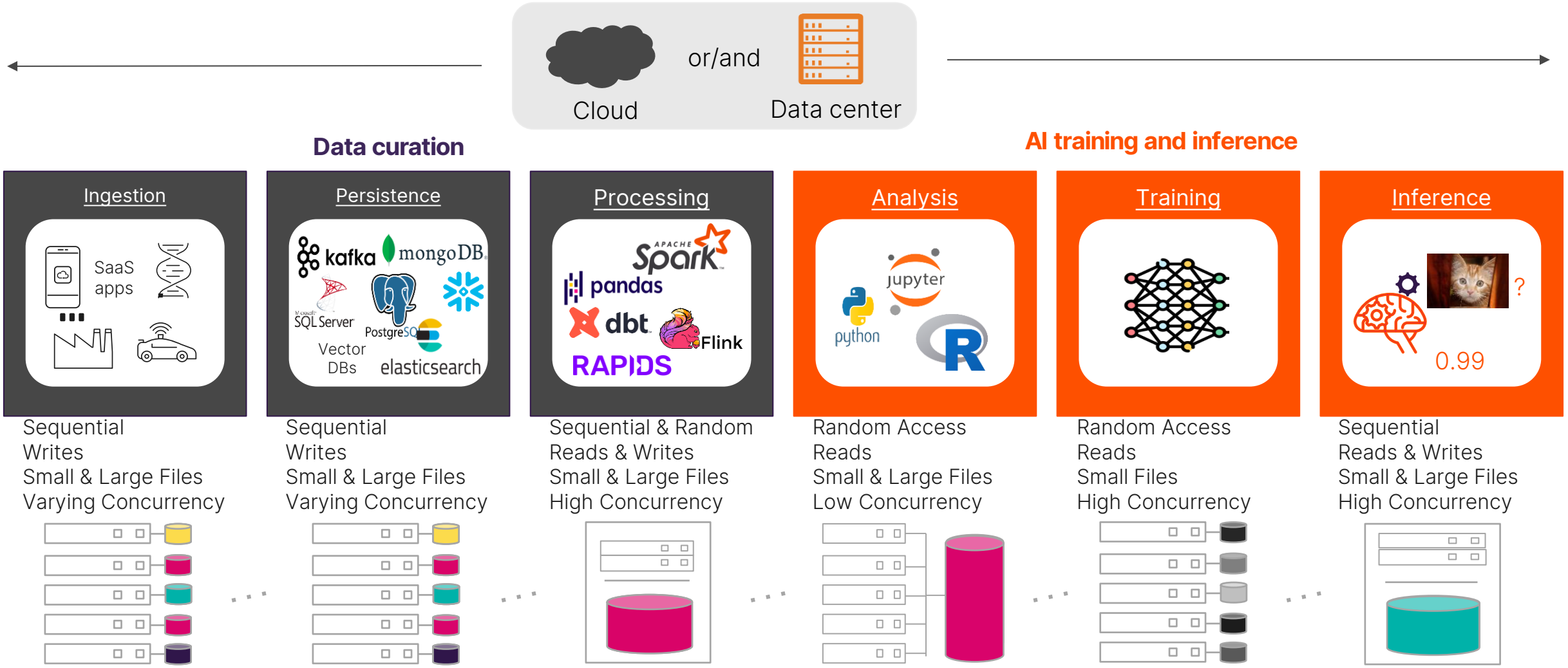
Cost Optimization

AI 트레이닝
비용 최적화

©2025 Pure Storage

1. AI Acceleration: AI 개발 도전과제

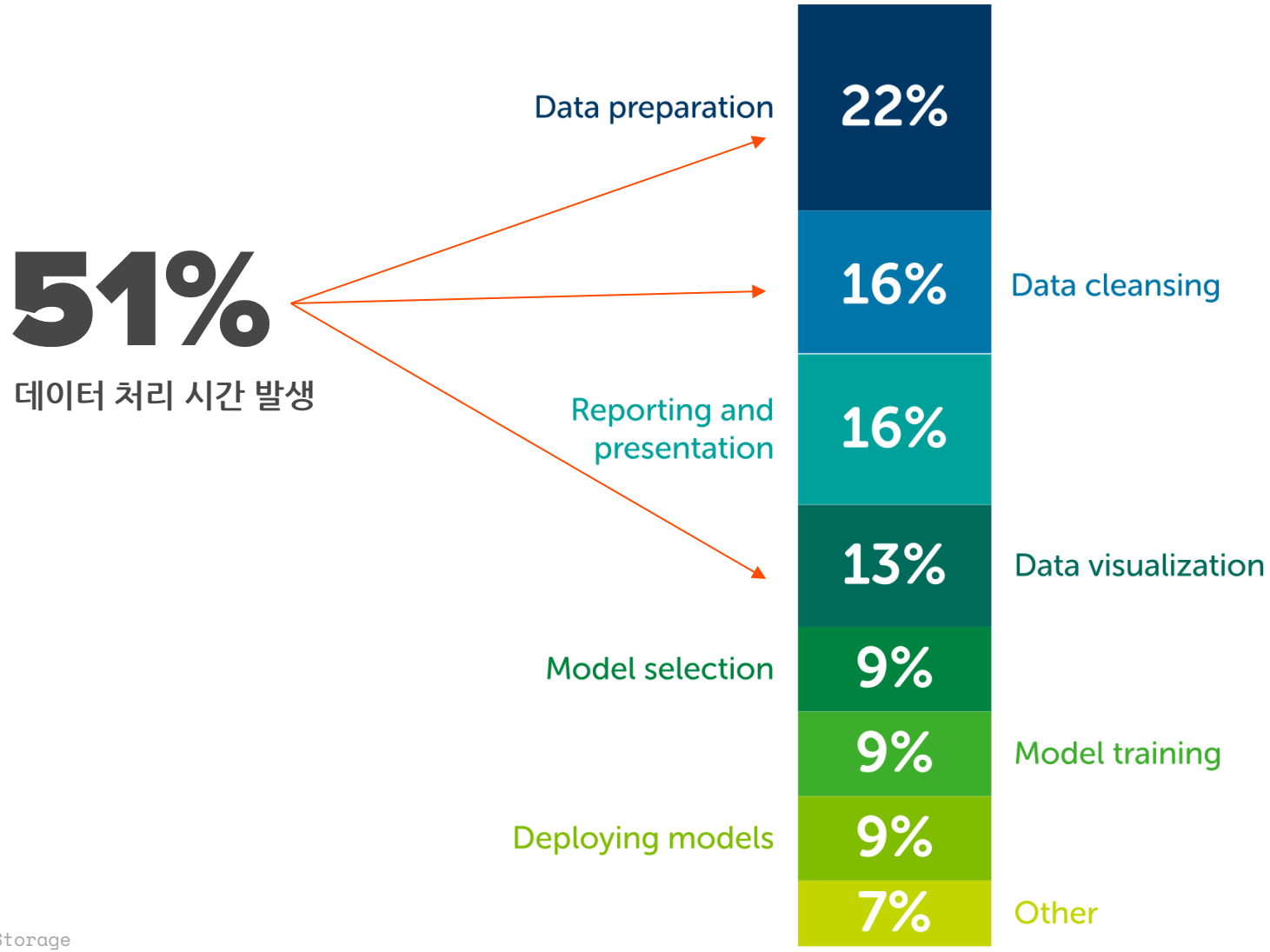
Training is just a small piece of a much larger & more complex picture



©2025 Pure Storage

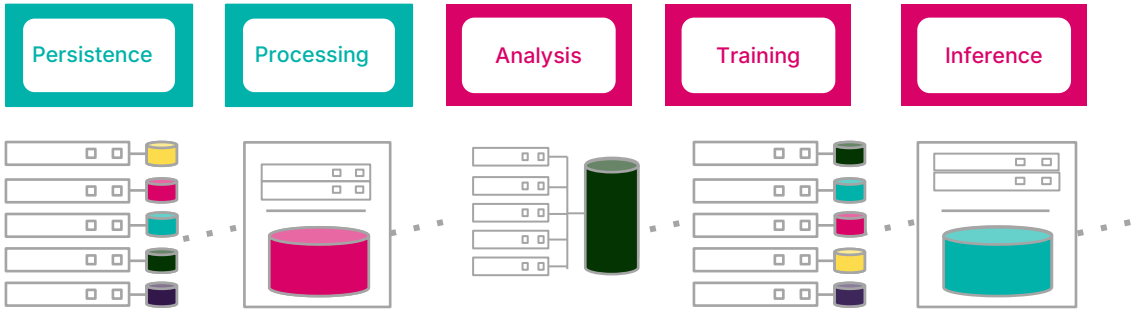
1. AI Acceleration: 비효율적인 데이터처리

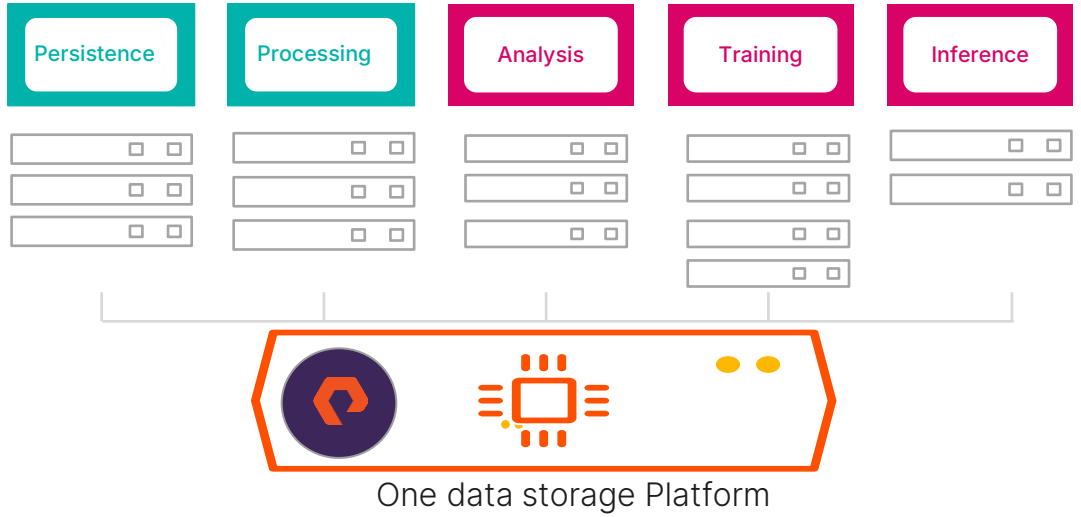
Anaconda "2022 State of Data Science"



©2025 Pure Storage

1. AI Acceleration: 데이터 통합





사일로 관리 어려움

데이터 타입에 따른 성능 편차

제한된 확장성

반복적인 데이터 복사

단일 관리 플랫폼 기반 간편한 운영

데이터타입에 상관없는 일관된 성능

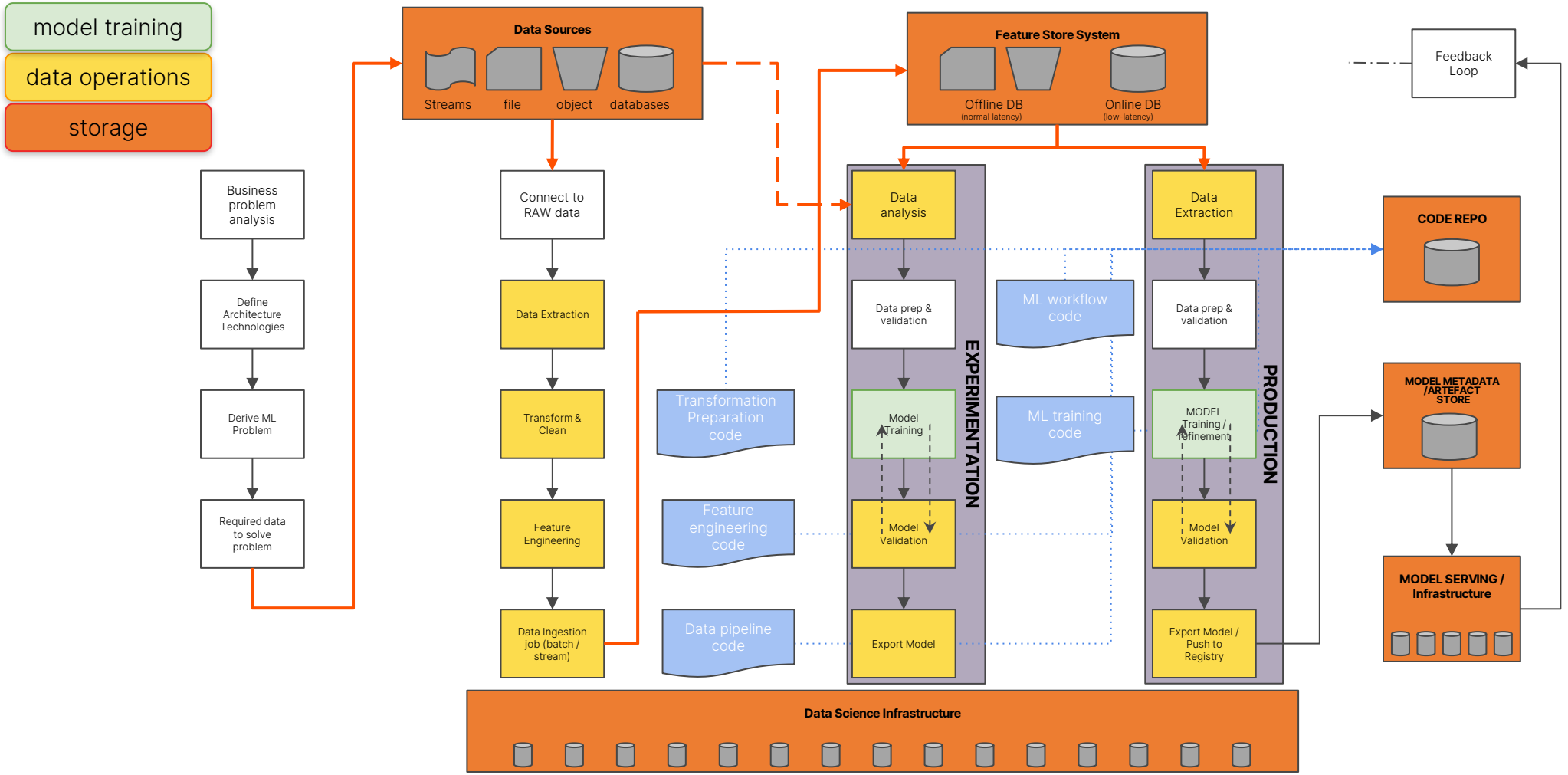
유연한 확장성

데이터 처리 시간 최소화

데이터 통합을 활용한 AI 가속화

1. AI Acceleration: 데이터 사일로 구성 예시

An real example of what goes into an AI Pipeline



©2025 Pure Storage

1. AI Acceleration: 데이터 통합 및 일관된 성능 보장

Train your model in days instead of months

any job | any protocol | any size | any object count | any processing type

Ingestion | Persistence | Processing | Training | Inference

AI workloads require multi-dimensional performance

예측 가능한 확장성

컴퓨팅과 스토리지 간
유연한 성능 확장

예측 가능한 성능

별도 튜닝 없이
일관된 성능 보장

FlashBlade 플랫폼

사용자 친화적 인프라

직관적이고 심플한
사용자 인터페이스

운영비용 절감

85% 개선된
에너지 효율성

©2025 Pure Storage

1. AI Acceleration: 데이터 통합 및 일관된 성능 보장

AI / HPC 전용 PURE 플랫폼

Enterprise AI and HPC
RAG / Inference / Data Lake / ML

AI Factories
Large Scale AI Training/Inference including Pre-processing, Training, Testing, Fine-tuning, and Deployment

- Enterprise AI/ML & HPC: Repository training, Archive, & Raw data
- Enterprise HPC and EDA: Active chip design projects
- Enterprise AI: Training and Inference
- Extreme Scale HPC Workloads
- Extreme Scale End to End AI workflows

TCO Optimized → Enterprise Ready High Performance → Extreme Performance and Capacity

FlashBlade//E
The All-Flash Unstructured Data Repository
Up to 90% Lower Operational Cost

FlashBlade//S
The Unified Fast File and Object Solution
Up to 500 GB/s Throughput

FlashBlade//EXA
The Extreme Scale Solution for Large HPC and AI Workloads
Up to 10+ TB/s Throughput, Exascale Capacity

©2025 Pure Storage

1. AI Acceleration: 데이터 처리 특화 기술

RapidFile Toolkit을 통한 데이터 전처리 작업 가속화

[RapidFile Toolkit]

관리자 생산성 획기적 증가

- 최대 200배 빠른 파일 처리 및 관리
- 파일 관리 및 분석 가속화

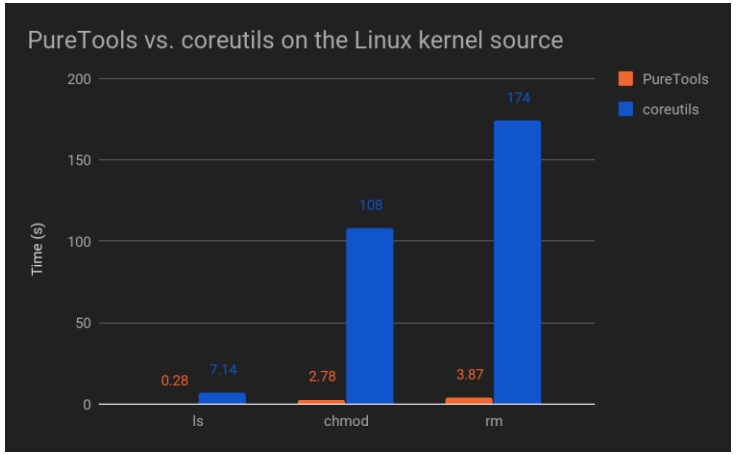
더 빨라진 데이터 이동 및 분석

- 스크래치 데이터 간 이동 속도 개선
- pfind 를 통한 데이터 검색 성능 개선

빠르고 간편해진 데이터 파이프라인 구축

- ptar 를 활용한 데이터 아카이빙 지원
- EDA, 유전체학, DevOps, HPC, 분석 및 Apache Spark, AI/ML 지원

데이터 큐레이션 시간 개선 및 AI 전용 파일 관리 기술



Linux 명령어	RFT 명령어	Example
ls	***pls***	pls -R -l /path/to/directory | pls -R --jsonlines /path/to/directory
find	***pfind***	pfind /path/to/files-to-delete -type f -atime +365 -print0 | prm -0
du	***pdu***	pdu -hs /path/to/directory
rm	***prm***	prm -r /path/to/dir-to-delete
chown	***pchown***	pchown -R NEWUSER:NEWGROUP /path/to/files-to-modify
chmod	***pchmod***	pchmod -R NEWPERMS /path/to/files-to-modify
cp	***pcopy***	pcopy -rfp /path/to/source/. /path/to/target
tar	***ptar***	ptar -xvf archive.tgz

©2025 Pure Storage

1. AI Acceleration: 데이터 처리 특화 기술

RapidFile Toolkit 실제 고객 POC 결과 사례

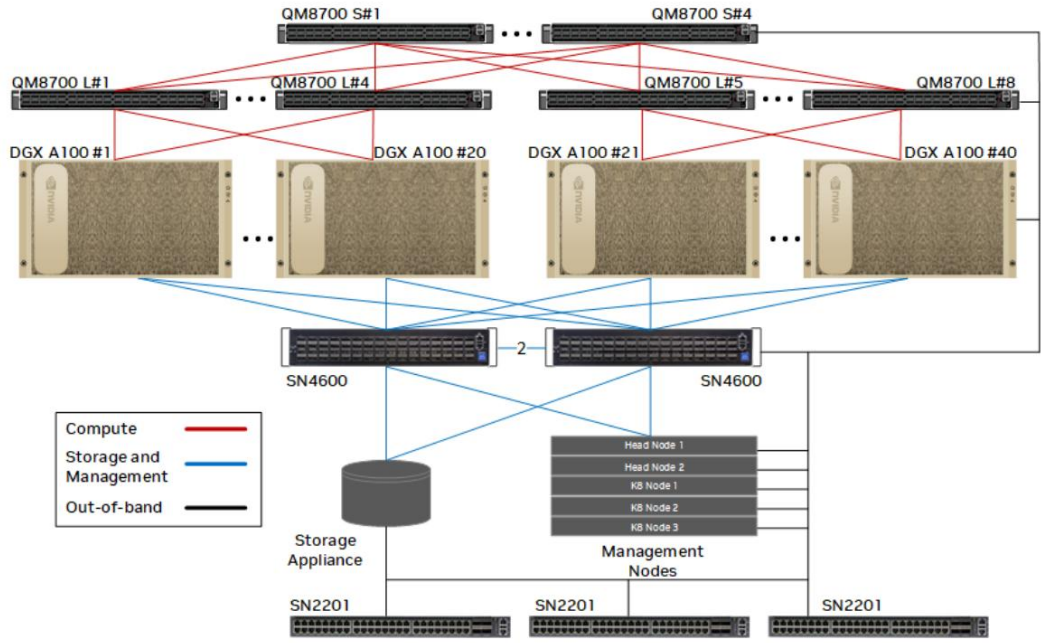
구분	Linux Command	RapidFile Tool-kit	비고
Command	find /pure/6Bil_files -type f |wc -l	**pfind /pure/6Bil_files -type f |wc -l**	일반 리눅스 명령어 **"find"** 와 RapidFile **"pfind"** 비교
소요시간	**13,635분 (9.4일)**	**182분 (약 3시간)**	***215 배 (21500%)의 성능 개선 확인***
수행 결과	Mon Jun 5 07:32:32 KST 2023 2195410539 real 13635m41.800s user 36m47.975s sys 817m12.997s Wed Jun 14 18:48:14 KST 2023	Wed Jun 14 19:00:01 KST 2023 6360373594 real 182m11.744s user 595m17.458s sys 93m34.643s Wed Jun 14 22:02:13 KST 2023	일반 리눅스 명령어 find 수행 시, 9일 경과 후에도 미완료 → 프로세스 강제 종료 **종료 시점 찾은 파일 갯수는 약 21.9억개**
Command	rm -rf /pure/6Bil_files	**prm -rf /pure/6Bil_files**	일반 리눅스 명령어 **"rm"** 과 RapidFile **"prm"** 비교
소요시간	**14,517분 (241.9시간/약 10일)**	**836분 (13.9시간)**	***121배(12100%) 이상의 성능 개선 확인***
	Mon Jun 19 09:17:58 KST 2023 ./rm.sh: line 2: 1114 Killed real 14517m36.288s user 13m28.835s sys 427m56.093s Thu Jun 29 11:15:34 KST 2023	Fri Jun 16 12:20:36 KST 2023 real 836m36.963s user 1781m31.602s sys 284m54.745s Sat Jun 17 02:17:13 KST 2023	일반 리눅스 명령어 rm 수행 시, 10일 경과 후에도 미완료 → 프로세스 강제 종료 **삭제 시간이 약 70일 소요될 것으로 예상**

©2025 Pure Storage

2. Disaggregated Architecture: 컴퓨트-스토리지 연결

컴퓨팅과 스토리지를 단일 네트워크로 구성?! → ***개별 네트워크 구성으로 성능 보장 및 장애 도메인 분리***

Figure 18. DGX BasePOD configurations with up to 40 systems—DGX A100 HDR



- **Disaggregated 구성**
- **스토리지 연결과 노드간 통신 네트워크 분리**
- **서비스 별 트래픽 분리 / 네트워크 혼잡 제어**
- **네트워크 장애 시 영향도 최소화**
- **높은 스토리지 사용율**

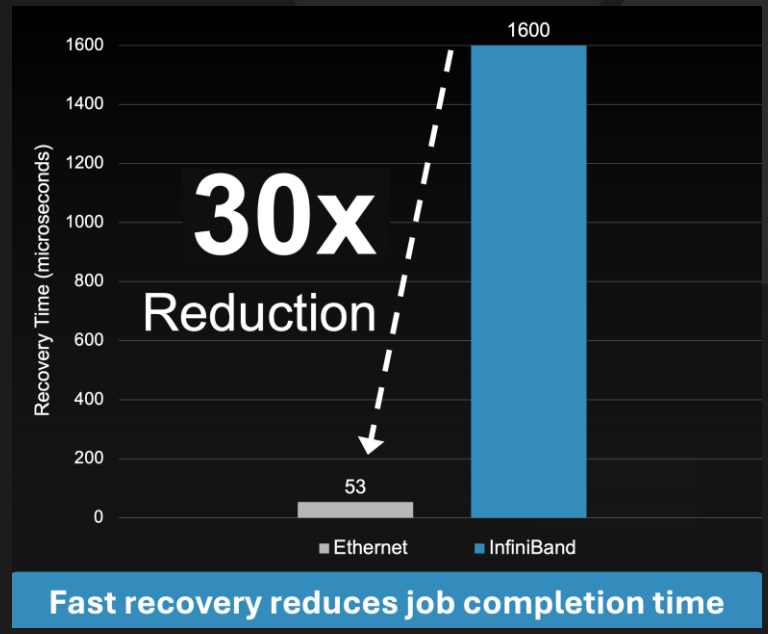
©2025 Pure Storage

2. Disaggregated Architecture : ETH vs IB

AI 워크로드에 대한 높은 성능 및 가용성 확인

AI 워크로드 수행 시, 인피니밴드 대비 10% 수행시간 개선 확인

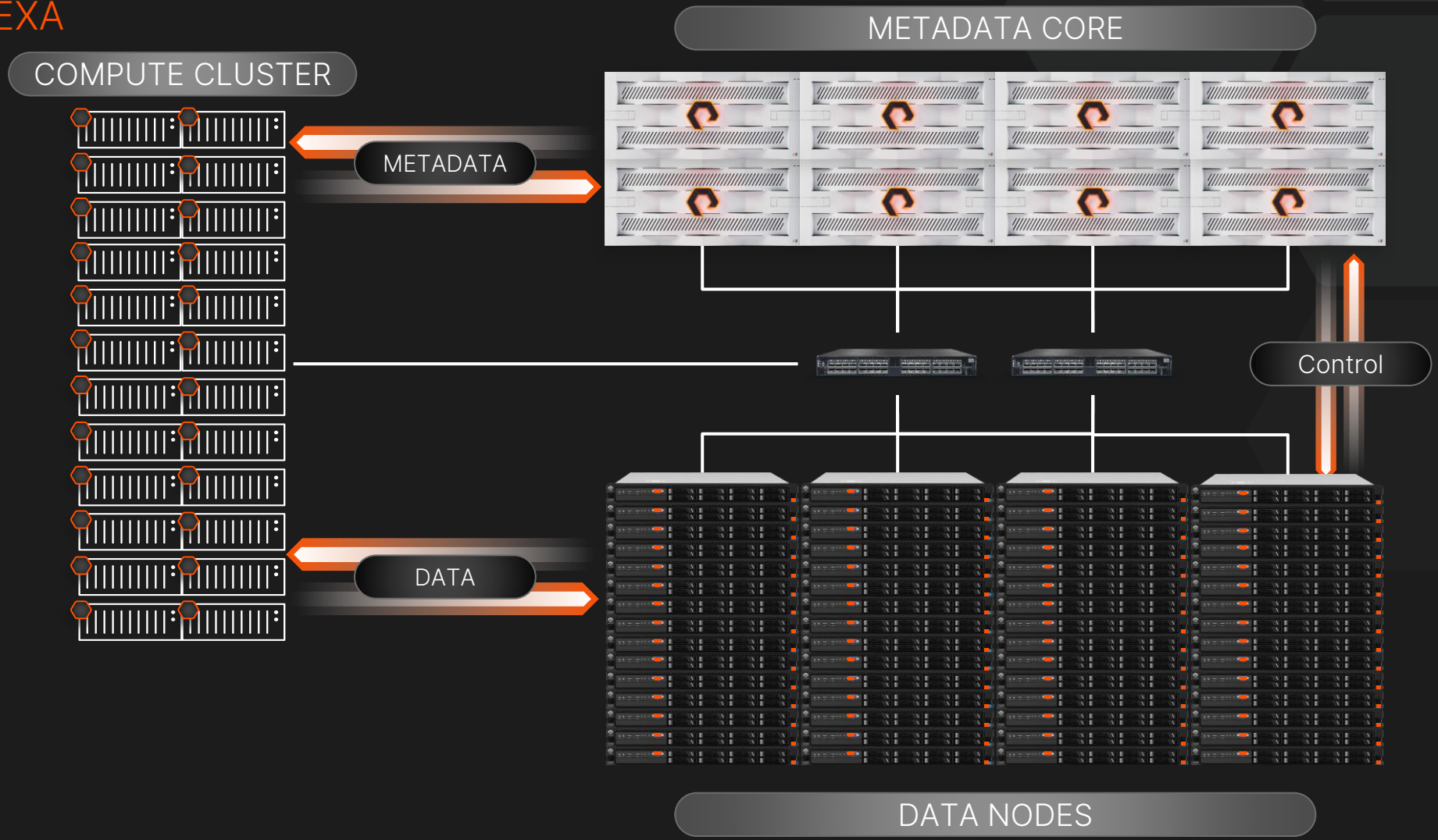
인피니밴드 대비 Failover 시간 30% 개선 확인



Fast recovery reduces job completion time

2. Disaggregated Architecture: 스토리지 확장성

FlashBlade//EXA



2. Engineered AI-Ready Infrastructure

FlashBlade

RAG Platforms with NVIDIA NIM, DGX and OVX

FlashBlade//S
+ NVIDIA DGX or OVX

Pure Validated RAG & Inference RAs

NVIDIA

GenAI Pod Turnkey, Full-Stack

NVIDIA, DGX , Supermicro (OVX, HGX)

Automated Deployment of NVIDIA NIM, LLM & Vector DBs

portworx

Cisco Validated Design

FlashStack for AI

GenAI Inferencing, MLOps & RAG

portworx

NVIDIA DGX BasePOD

AIRI® - AI-Ready Infrastructure

portworx

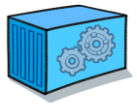
NVIDIA DGX SuperPOD

Turnkey Enterprise AI

Certified *Ethernet* Solution

©2025 Pure Storage

3. Deployment Consideration



Why Containers for AI (and everything else)?

AI 파이프라인 패키징, 간편한 배포 → 신속한 개발, 테스트 및 프로덕션 배포

- Portability
- Scalability
- Performant
- Isolation
- Efficiency
- Reproducibility
- DevOps integration
- Security
- Consistent environments
- Resource management

Why Kubernetes?

오픈소스 기반 컨테이너 관리 플랫폼 / 컨테이너 관리 표준 (De facto standard)

- Automated deployment and scaling
- Resource management
- Load balancing
- Auto restart of failed containers
- CI/CD integrations
- Monitoring and logging
- On prem, cloud, or hybrid support

개발 환경에 대한 최선의 선택, ***프로덕션을 위한 최선의 선택?***

©2025 Pure Storage

3. Deployment Consideration

- 컨테이너용 영구 스토리지 기능 및 Stateful 에 대한 부족함
- 데이터 복제 / 재해복구 기능 부족
- 멀티 클라우드 및 하이브리드 클라우드 환경에서의 복잡한 데이터 관리
- 암호화 및 보안 액세스 제어와 같은 데이터 보안 기능 미제공
- 고급 백업 및 복원 기능 부족

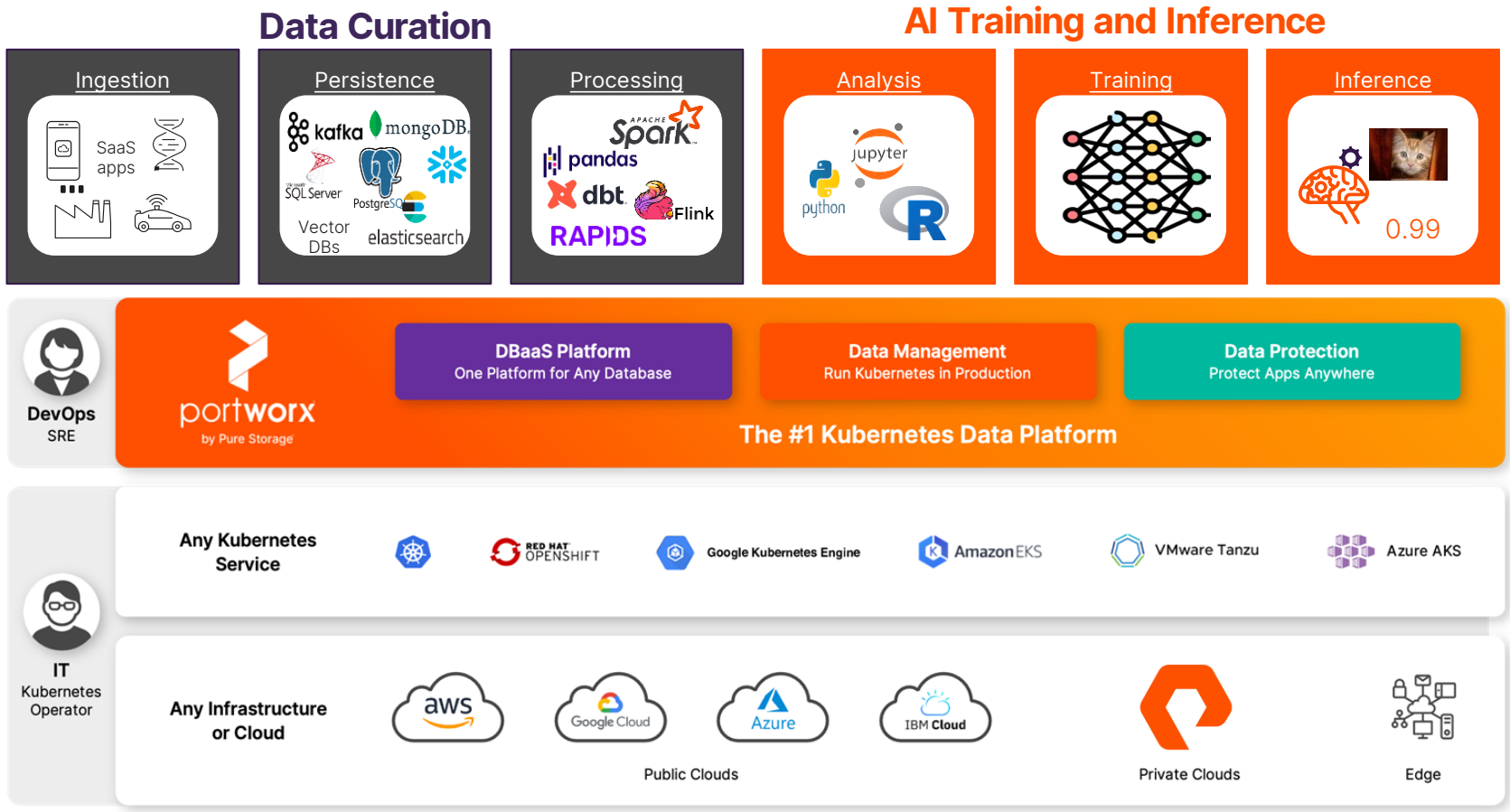
©2025 Pure Storage

3. Deployment Consideration

Portworx 를 통한 컨테이너 데이터 관리 및 엔터프라이즈 기반 프로덕션 배포

Cloud-native 기반의 성공적인 AI/ML 워크플로우 구축

- 모델 및 실험에 대한 파이프라인 가속화
- K8s 기반 컨테이너 동적 확장 지원
- 개발자를 위한 셀프서비스 기반 DB-as-a-Service 제공



©2025 Pure Storage

4. Cost Optimization : AI 비용 구조

학습 비용

- HW, SW 비용 + 애플리케이션 개발 비용
- 모델 복잡도 및 개발 사이클에 의존
- 단기적인 지출

→ ***개발 주기 단축 시 절감***

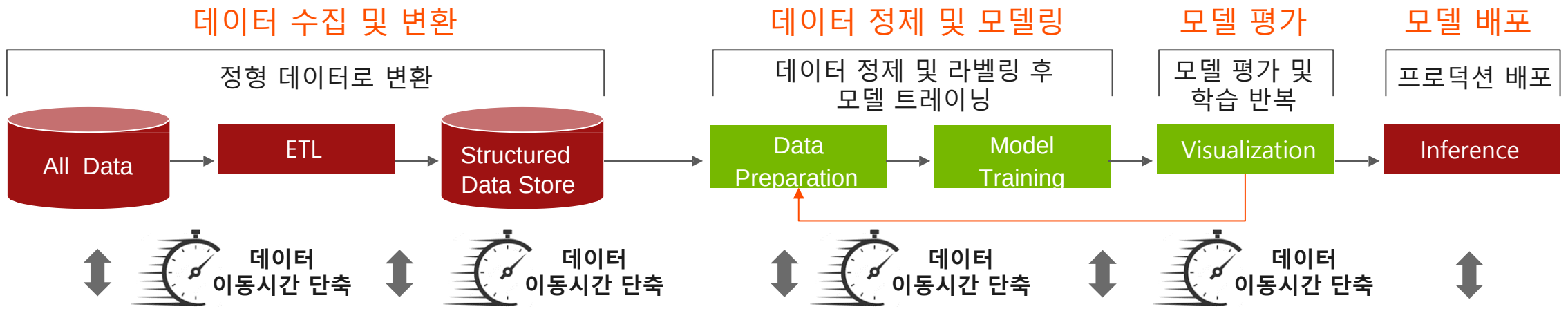
추론 비용

- 하드웨어, 소프트웨어 유지 비용 (전력, 상면, 운영 인건비 등)
- 모델 복잡도 및 활용 빈도에 의존
- 장기적인 지출

→ ***모델 최적화/간소화***
→ ***하드웨어 효율화 시 절감***

©2025 Pure Storage

4. Cost Optimization : AI 데이터 통합



portworx by Pure Storage

FlashBlade //S

One Data Platform

©2025 Pure Storage

4. Cost Optimization : 운영비용 절감

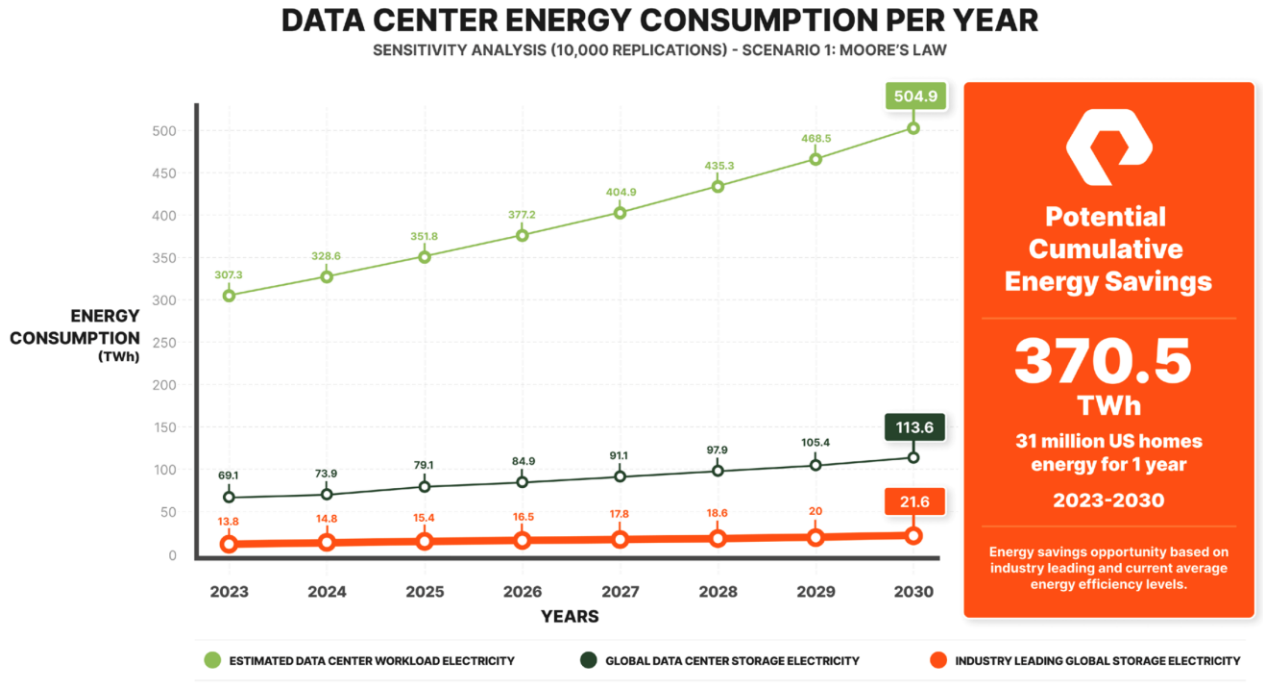
- 낮은 디스크 장애율 기반 7X 높은 운영 신뢰성 보장
- 에너지 소비량 85% 절감 및 데이터센터 상면 80% 절감

• 최근 5년간 국내 3개 고객사 DFM 디스크 장애 건 → 3건

Account	Num. of arrays	Num. of DFMs	Total returned DFM for 5y.	Annual Return Rates(%)
Company-a	7	264	0	0%
Company-b	87	2,281	3	0.026%
Company-c	21	338	0	0%
Total	**94**	**2,883**	**3**	**0.020%**

$$\text{ARR(Annual Return Rate)} = \left(\frac{\text{DFM Return number per 1y}}{\text{Total DFM number}} \times 100 \right) / 5(y)$$

0.0002 DFM failed per 1 year



©2025 Pure Storage

4. Cost Optimization : 서비스형 과금

Think **consumption**, not capacity. **Data**, not storage.

Services		For AI	Block and File				File and Object			UDR	
Tiers		AI	Ultra	Premium	Performance	Capacity	Premium	Performance	Standard	Block & File	File & Object
Example Workloads		Training, inference, RAG, machine learning, computer vision, deep learning, LLMs	In memory database, Oracle, SAP HANA, Tier 0 database	VDI, production database	Tier 1 workloads, VSI (VMs), production database, containers, DR to cloud	IoT, backup & retention, Tier 2 database & VMs, test/dev, snapshots	Imaging & rich media, analytics	IoT data, software development, S3 enabled apps	Rapid restore, artifact repository	Information preservation, long term backup & retention, regulatory/ compliance archiving, spinning disk replacement	Picture archiving & communication systems (PACS), surveillance video footage, exploratory data analysis
SLA	Minimum Performance	**30 GB/s** provisioned	**48 MB/s** per TiB EUC	**24 MB/s** per TiB EUC	**8 MB/s** per TiB EUC	**1.6 MB/s** per TiB EUC	**50 MB/s** per TiB EUC	**20 MB/s** per TiB EUC	**10 MB/s** per TiB EUC	**0.5 MB/s** per TiB EUC	**1 MB/s** per TiB EUC
	Max Energy Consumption	**100 W** / GB/s **7 W** / TiB EUC	**7 Watts** per TiB EUC	**4 Watts** per TiB EUC	**3 Watts** per TiB EUC	**2 Watts** per TiB EUC	**4 Watts** per TiB EUC	**4 Watts** per TiB EUC	**4 Watts** per TiB EUC	**3 Watts** per TiB EUC	**2 Watts** per TiB EUC
	Minimum Availability	**99.9999%** (including planned maintenance)									
	Minimum Buffer Capacity	**25%** (at all times)									
Minimum Commit		**30 GB/s** of bandwidth	**50 TiB** (EUC)			**200 TiB** (EUC)	**100 TiB** (EUC)	**100 TiB** (EUC)	**100 TiB** (EUC)	**500 TiB** (EUC)	**1500 TiB** (EUC)
Minimum Term		12 months								36 months	

©2025 Pure Storage

도입 사례 #1 - LandingAI & CoreWeave

June 5, 2024

Pure Storage, LandingAI 에 전략적 투자해 기업 AI 비전 모델 발전 도모

LandingAI의 주력 제품인 LandingLens™는 사용자가 컴퓨터 비전 솔루션을 빠르고 쉽게 구축, 반복 및 배포할 수 있도록 지원하는 컴퓨터 비전 클라우드 플랫폼입니다.

LandingAI는 또한 도메인별 대형 비전 모델(LVM) 분야의 선구자입니다. LVM은 대규모 시각 데이터 처리 및 이해 능력을 향상시켜 정교한 시각 AI 도구의 접근성과 효율성을 높여줍니다.

November 13, 2024

Pure Storage, CoreWeave와 전략적 투자 및 기술 파트너십 체결로 대규모 AI 클라우드 서비스 혁신 가속화

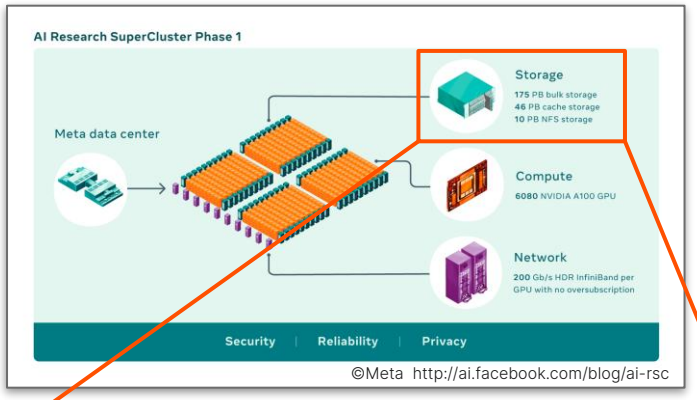
AI 하이퍼스케일러인 코어위브는 차세대 AI를 지원하는 최첨단 소프트웨어로 구성된 클라우드 플랫폼 기업입니다. 기업과 선도적인 AI 연구소에 컴퓨팅 가속화를 위한 클라우드 솔루션을 제공하며, 이를 위해 2017년부터 미국과 유럽 전역에서 데이터센터를 운영하고 있습니다.

코어위브는 타임지 선정 가장 영향력 있는 100대 기업 중 하나로 선정되었으며, 2024년에는 포브스 클라우드 100대 기업에도 이름을 올렸습니다.

"코어위브의 미션은 최고의 확장성과 성능을 갖춘 최첨단 AI 환경을 제공하는 것입니다. Pure Storage와의 파트너십을 통해 고객은 ***특정 AI 요구 사항에 맞는 스토리지 솔루션을 유연하게 선택***할 수 있습니다. 이번 협력을 통해 당사의 클라우드 플랫폼을 신뢰하는 고객에게 고속 성능, 안정성, 그리고 유연성을 제공하겠다는 약속을 더욱 공고히 할 수 있게 되었습니다."

CoreWeave 최고 전략 책임자(CSO)
브라이언 벤투로(Brian Venturo)

도입 사례 #2 - Meta

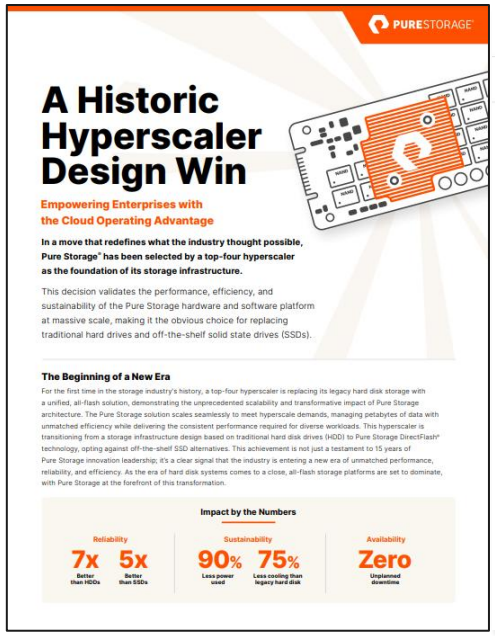


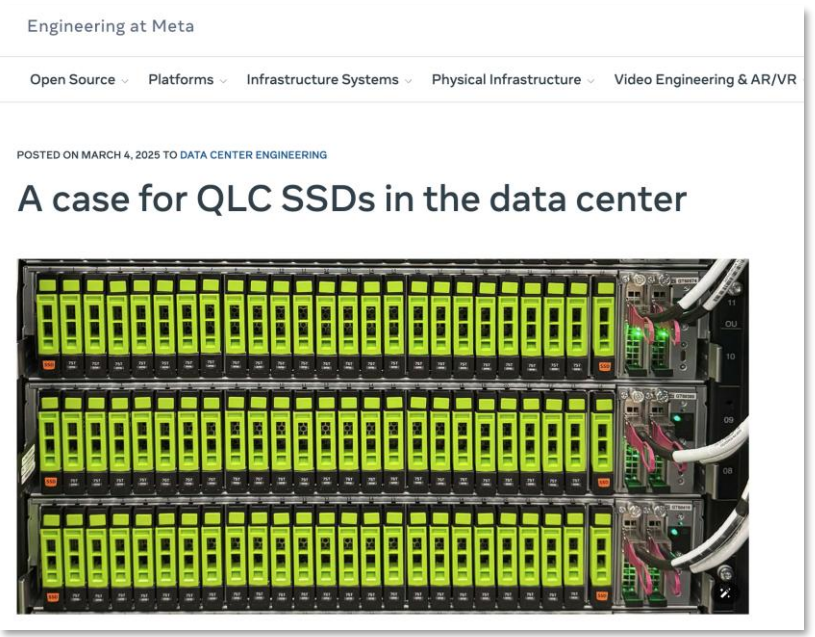
FlashArray //C

- 675PB of usable storage, fronted by a cache-tier
- Providing bulk storage for production model training
- Optimal capacity footprint (250TB/RU, 1.2W/TB)
- All-Flash Capacity tier (~10GB/s/PB)
- Denser and more power and cost efficient that HDD

FlashBlade® //FB

- 15PB of NFS Storage (3 clusters)
- Home dirs and working space for AI researchers
- Optimal Performance (198TB/RU, 2.25W/TB)
- High Throughput File & Object (~20GB/s/PB)





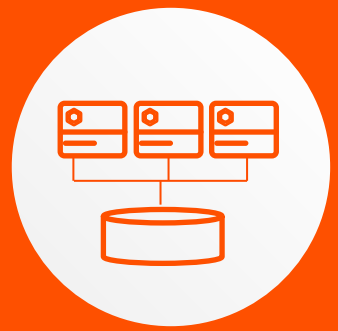
- **데이터 증가와 전력 효율성 향상에 대한 요구**는 혁신적인 스토리지 솔루션으로 연결
- HDD는 밀도는 증가했지만 성능은 그렇지 않았으며, **TLC 플래시는 확장에 제약이 되는 가격**
- **QLC 기술은 HDD와 TLC SSD 사이의 중간 계층을 형성하여 이러한 과제를 해결**하며, QLC는 기존 TLC SSD보다 더 높은 밀도, 향상된 전력 효율성, 그리고 더 나은 비용을 제공.
- **2026년 두 자릿수 엑사바이트 규모 달성 예상**

https://engineering.fb.com/2025/03/04/data-center-engineering/a-case-for-qlc-ssds-in-the-data-center/

요약: AI-Ready Infrastructure 기반의 AI 혁신

AI Acceleration

데이터 통합
일관된 성능



Dissagregated & Engineered

NVIDIA 인증
유연한 확장성 보장
이더넷기반 유연한 배포

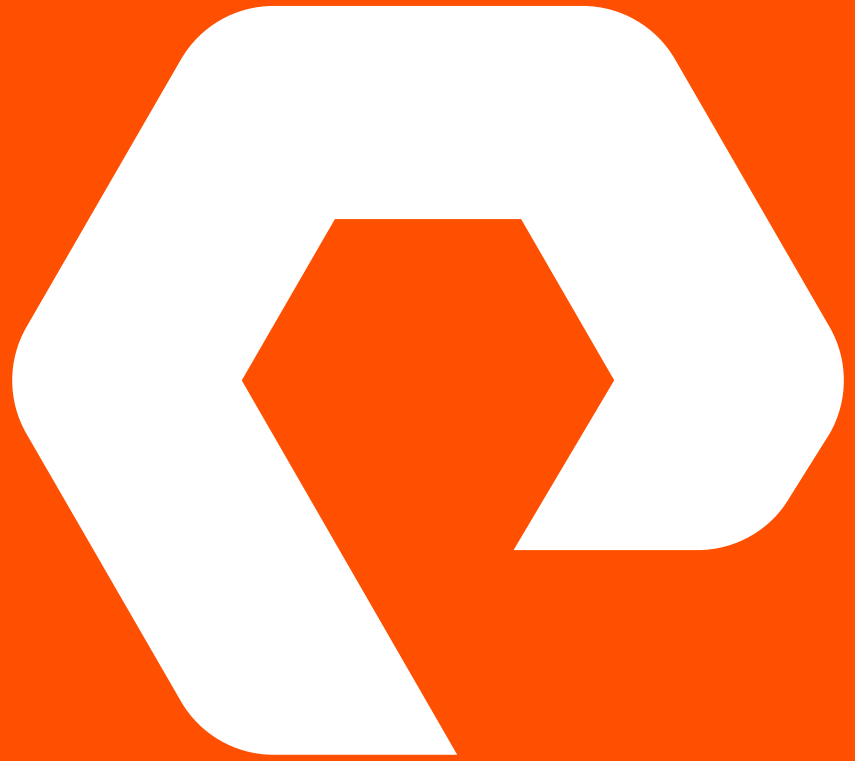
Deployment Consideration

Portworx 기반
컨테이너 데이터 관리

Cost Optimization

시간 단축 - 데이터 통합
운영비용 절감
구독 기반 인프라 활용

©2025 Pure Storage



Uncomplicate Data Storage, Forever