SNIA COMPUTE + MEMORY + STORAGE SUMMIT

*Architectures, Solutions, and Community*
*VIRTUAL EVENT, APRIL 11-12, 2023*

# Efficiency of data centric computing

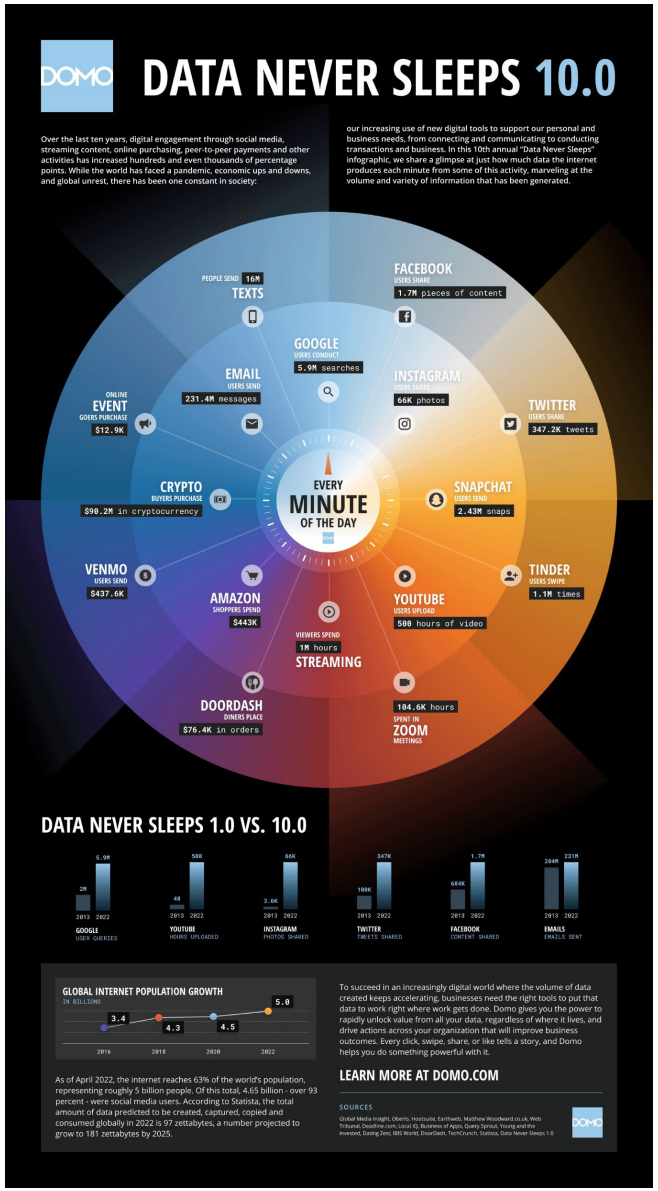
Presented by Steven Yuan
Founder & CEO of StorageX.ai

# Outline

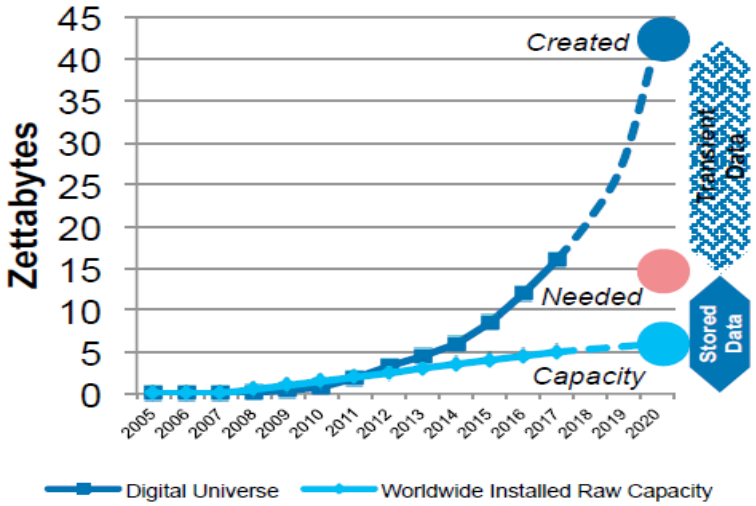
- ‘Big and Fast’ Data Challenges
  - Emergent Abilities of Large Language Models
- Optimize compute? Need to optimize data movements first
- Where Different Compute Resources Fit?
  - Dumbbell effect
  - Disaggregated systems
- Smart Data Lake
- Conclusion

# ’Big and Fast’ Data Demands New Compute + Storage Architecture

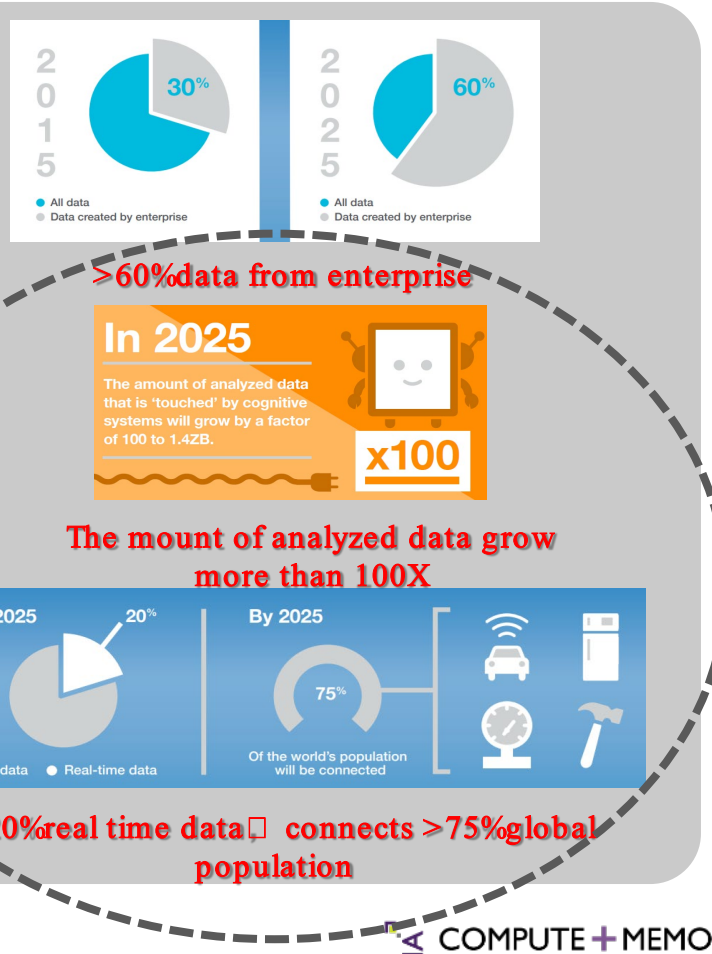



In 2020, we generated more than 40ZB of data



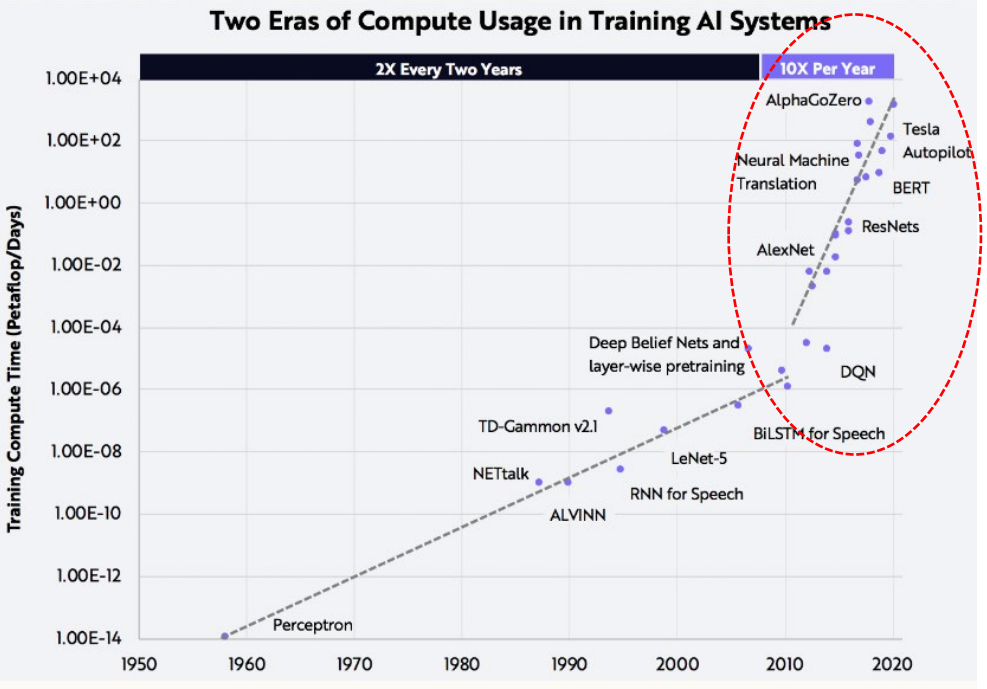

## 2022 data in every minute

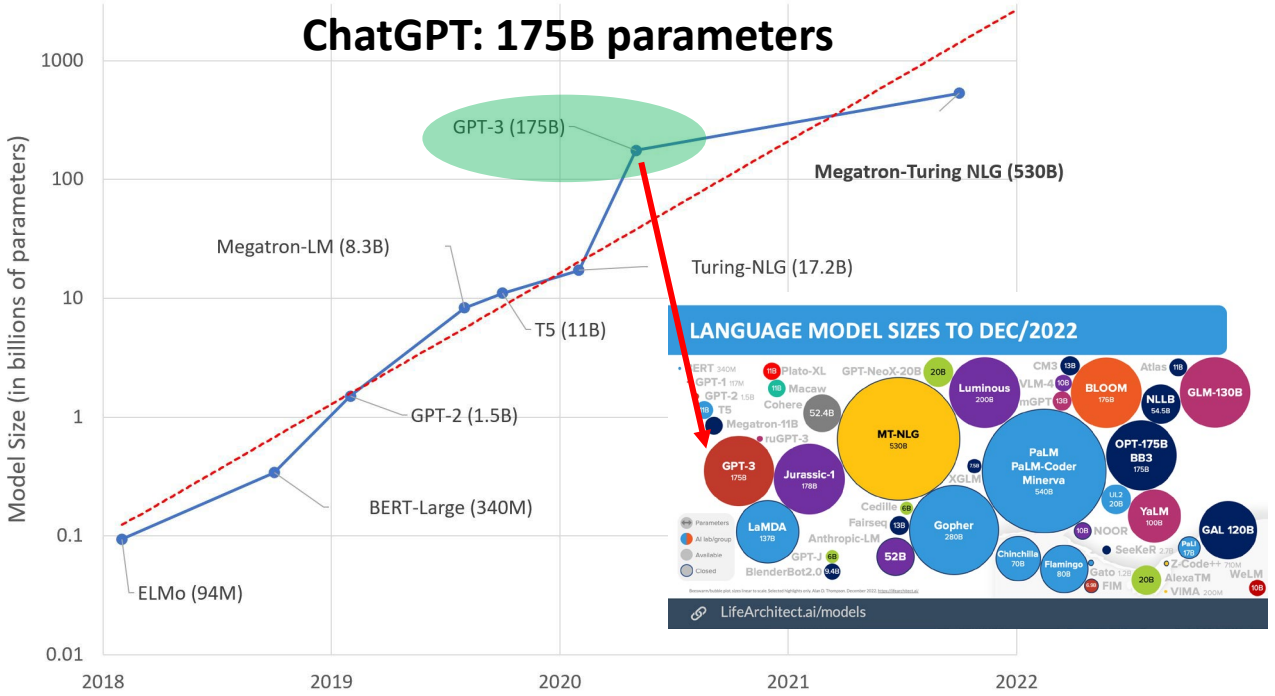
- **Google:** 5.9M searches
- **YouTube:** 500 hours of video upload
- **Steaming Video:** 1M hours
- **Facebook:** 1.7M of content shared
- **Zoom:** 105K hours of meeting

# Large models grows faster
# Balance of Compute & Data are becoming more predominant issue



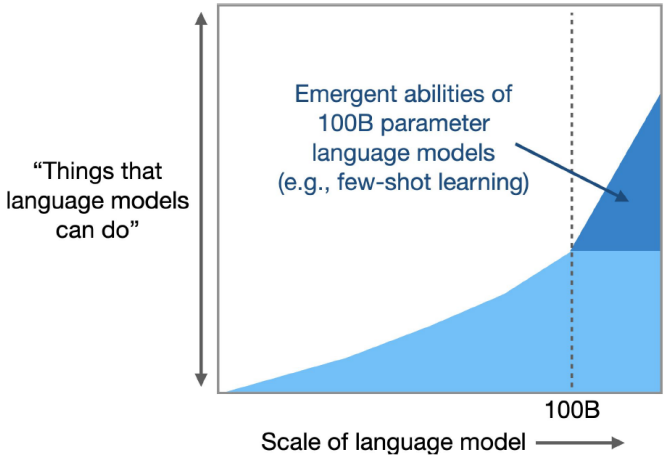



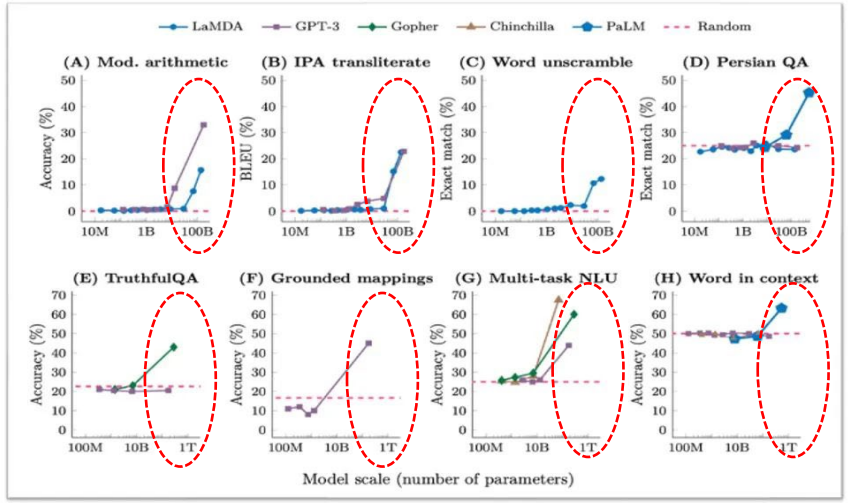

**Model size:**

**In the past, 2X per year (about same as Moore's law) → Now, 10X every year**

# Emergent Abilities of Large Language Models



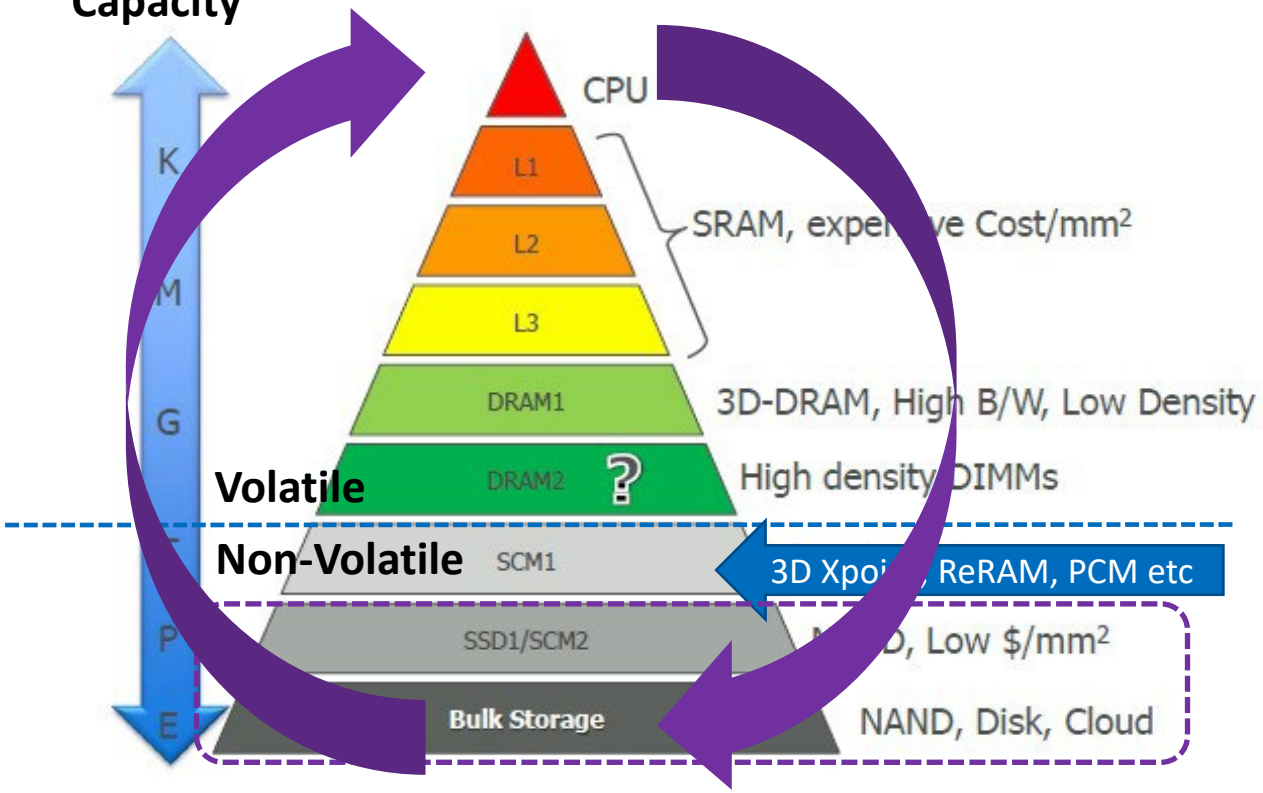




- **Compute, Data, and Neural Networks**. As hardware improved, it became possible to train neural networks that were very deep for the first time.
- **Better compute** enabled bigger models trained for longer, and **better storage enabled learning from more data**;
- **"Scaling unlocks emergent abilities in language models,"** Google researcher said a LLM technique called **chain-of-thought (COT)** prompting will bend the performance curve upward.
- While model size is over 100B, we're seeing unexpected **"emergent" capabilities coming out of their super-sized language models that is not presented in smaller models.**

# Optimize compute? Need to optimize data movements first

## Data Hierarchy




**Large scale of data@ PB & EB level, perfect area for data analytics**

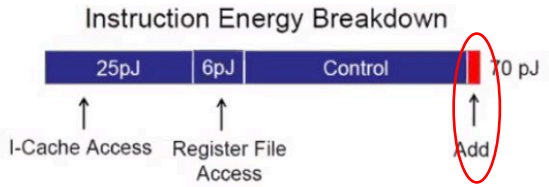
## Data Access and Data Movement Dominate Power Consumption

Rough Energy Numbers (45nm)

| Integer | |
|---|---|
| Add | |
| 8 bit | 0.03pJ |
| 32 bit | 0.1pJ |
| Mult | |
| 8 bit | 0.2pJ |
| 32 bit | 3 pJ |

| FP | |
|---|---|
| FAdd | |
| 16 bit | 0.4pJ |
| 32 bit | 0.9pJ |
| FMult | |
| 16 bit | 1pJ |
| 32 bit | 4pJ |

| Memory | |
|---|---|
| Cache | (64bit) |
| 8KB | 10pJ |
| 32KB | 20pJ |
| 1MB | 100pJ |
| DRAM | 1.3-2.6nJ |




Source: Mark Horowitz, "Computing's energy problem (and what we can do about it)," ISSCC 2014.

- Minimize data movement
- Reuse data as much as possible
- Rethink architectures

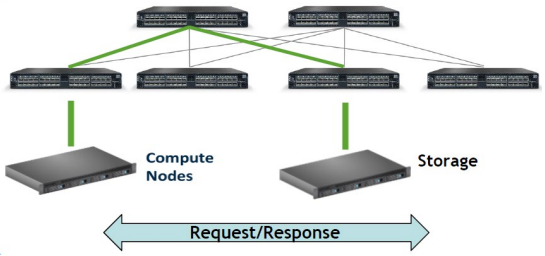
**Actual consumption of compute power only occupy small % of total energy consumed**

**0.03 – 0.9pJ of 70pJ → 0.04% to 1.3%**

# Where the Different Compute Resources Fit?

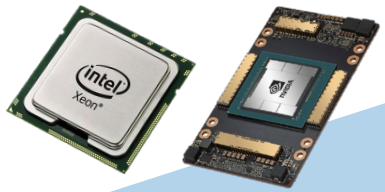
## Data/Compute Ratio




Latency is critical!

### Compute Node

- CPU
- GPU

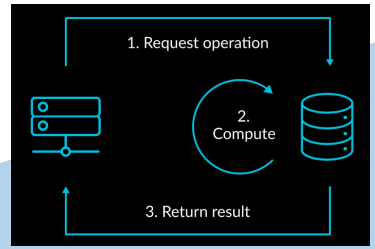



### In-network Computing

IPU/Smart NIC


- 100Gb/s -> 200Gb/s -> 400Gb/s
- Infrastructure
- Networking
- Security
- Orchestration

### Computational Storage (CS)




- Near Data Processing
- Data Acceleration
- Federated data processing
- AI/ML near the data

Fast Data

BIG DATA

Compute Intensive Workload

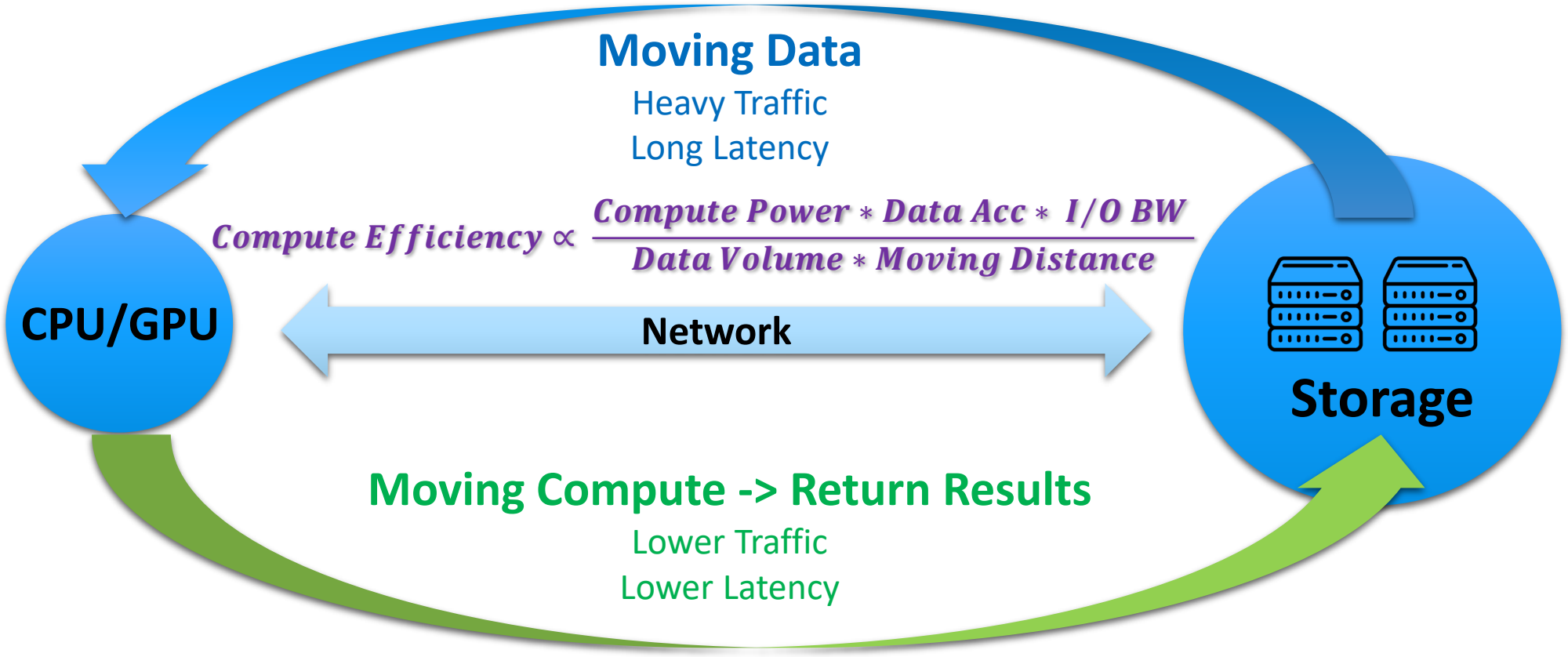
Small but Fast Data

Big Data but Need of Speed

Data Intensive Workload

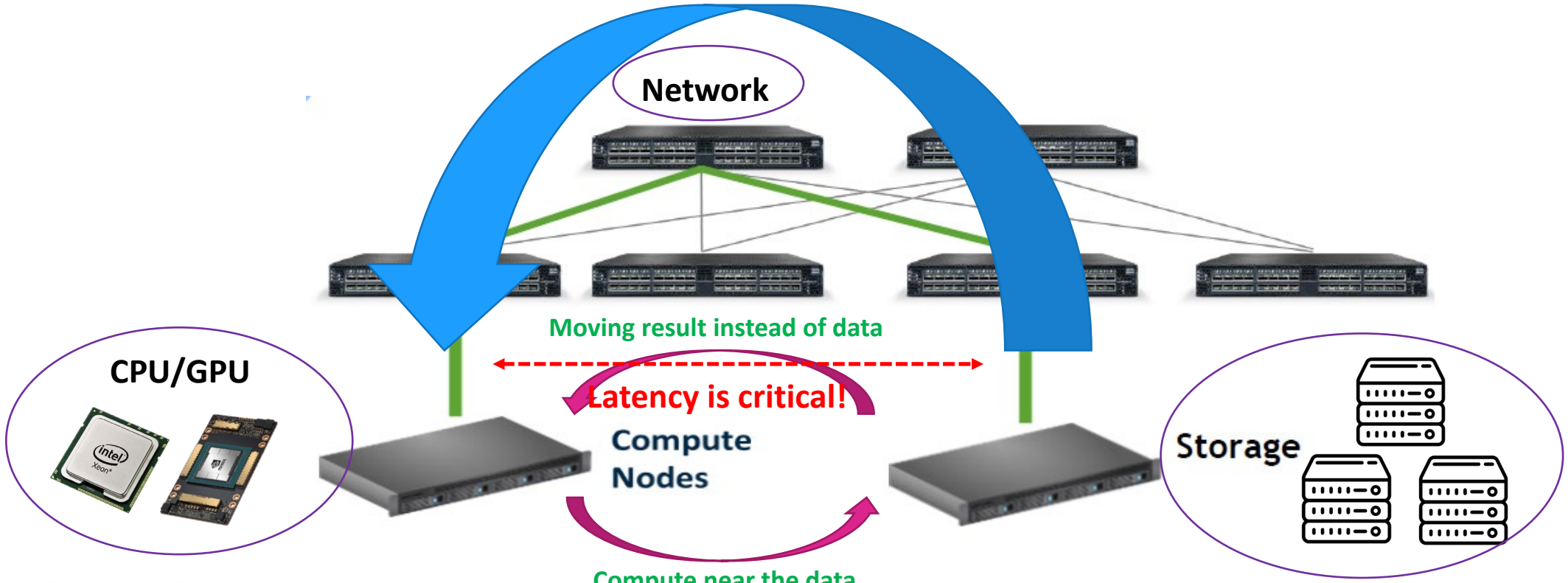
 

# The ‘Dumbbell effect’ causing high compute cost
# **Data Centric Computing** is a key solution

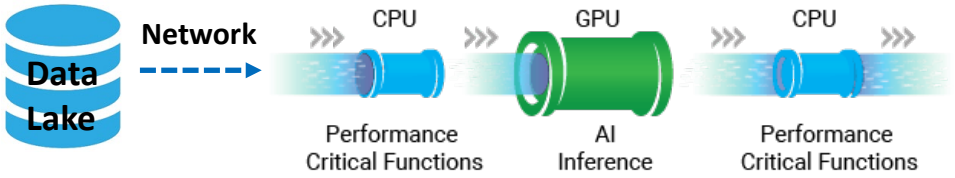


**Von Neumann's Architecture requires**
**large amount of data movements**
**high cost of network bandwidth, time & power consumption.**

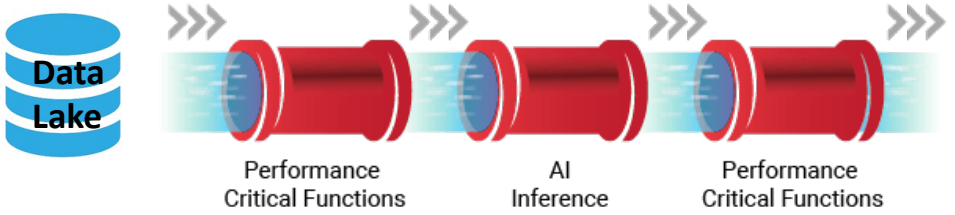



**Data Lake, Network, CPU, GPU**
**– Mismatched Throughput**



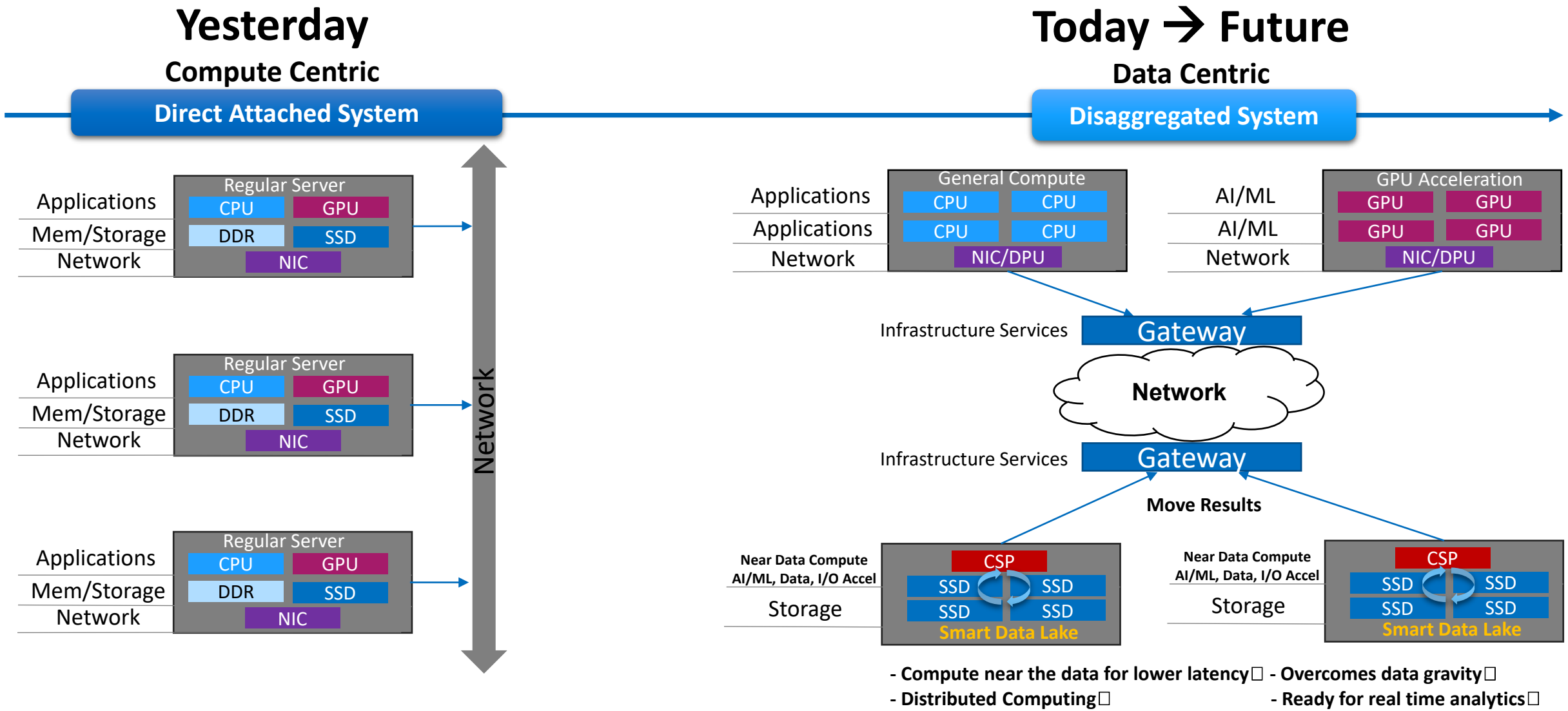

**Compute near the data**
**Reduces data movements, latency &**
**TCO drastically**

**Streamlined Throughput**



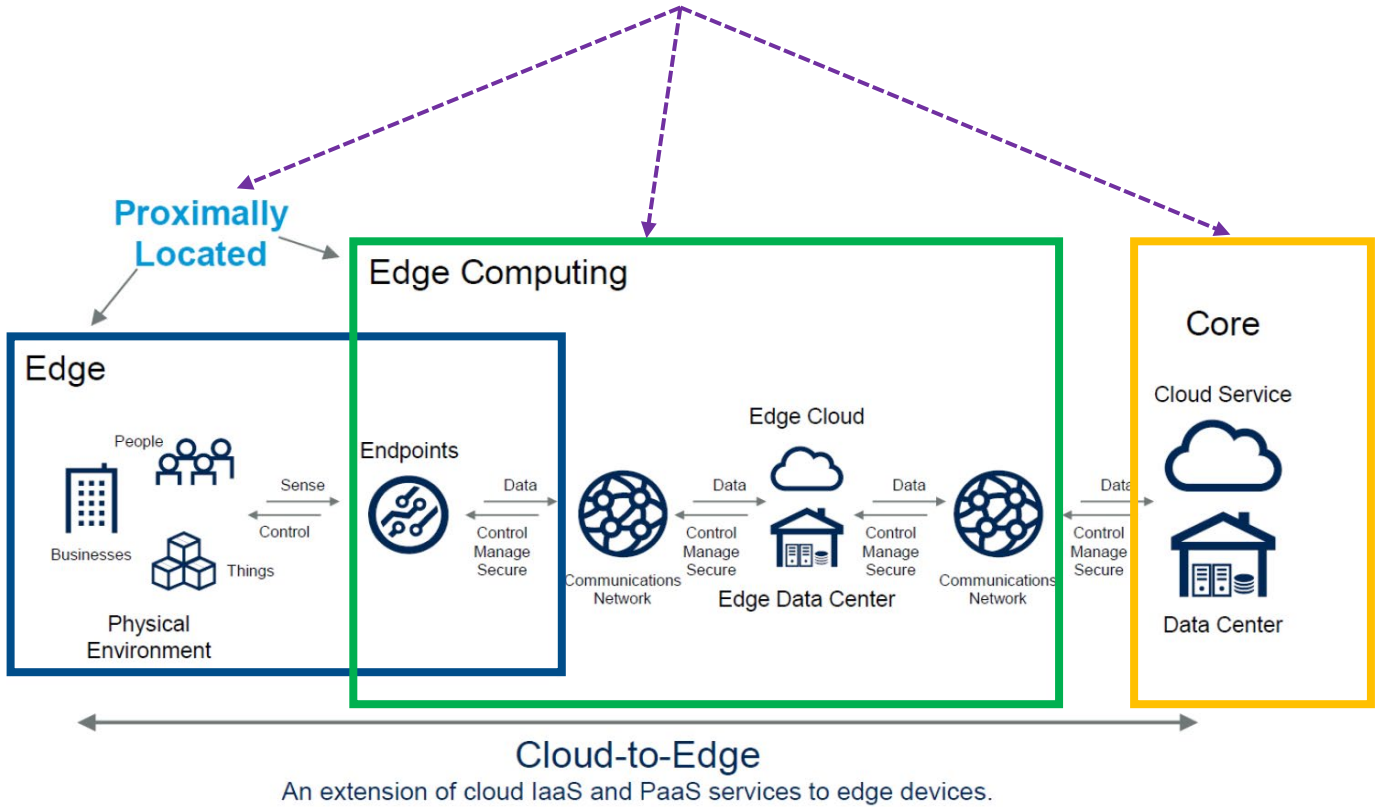
# Data center is moving to disaggregated systems




- Compute near the data for lower latency
- Overcomes data gravity
- Distributed Computing
- Ready for real time analytics

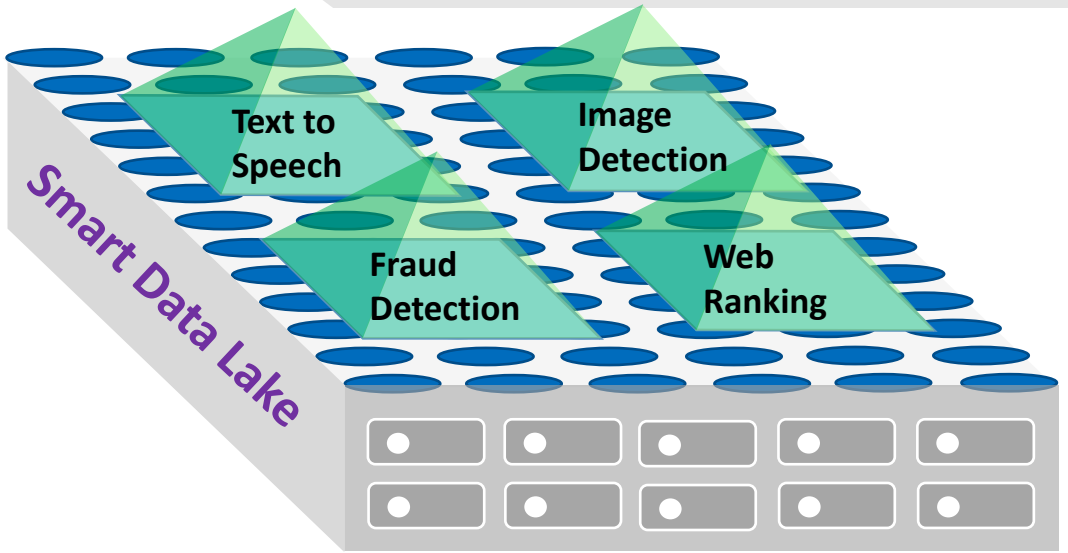
 

# Smart Data Lake




***Usages:***

- Compute Near Data
- Data Acceleration
- Real Time Analytics
- Flexible Applications

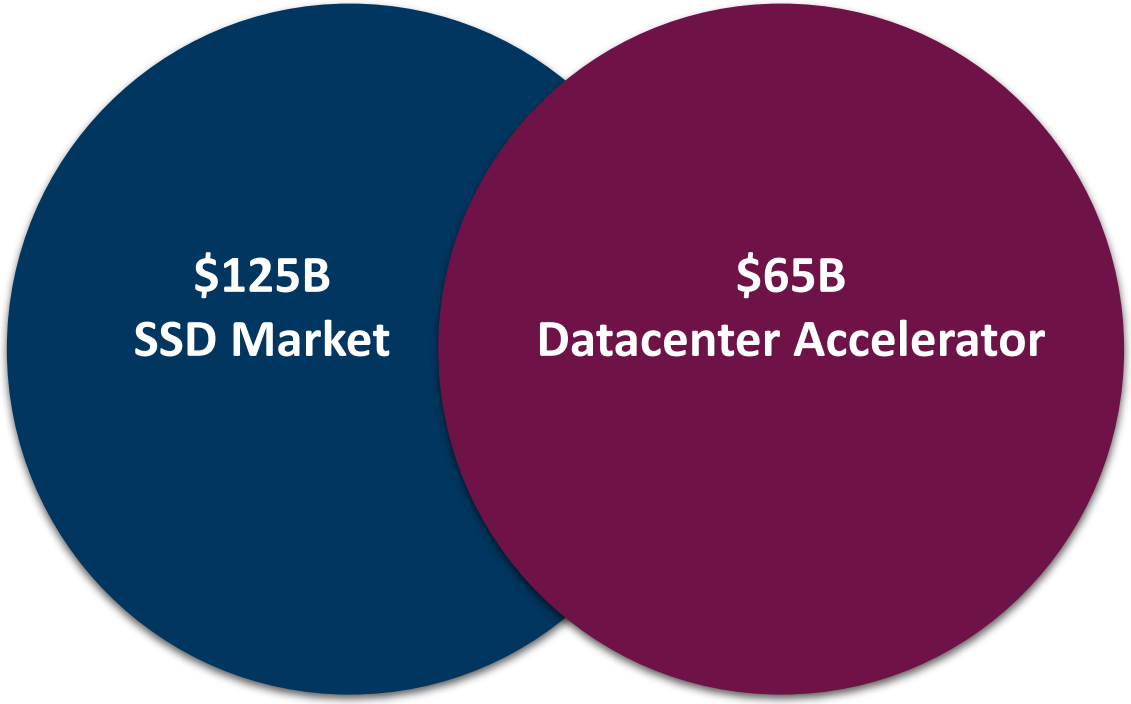



**Edge** → **Edge Cloud** → **Hyperscale**

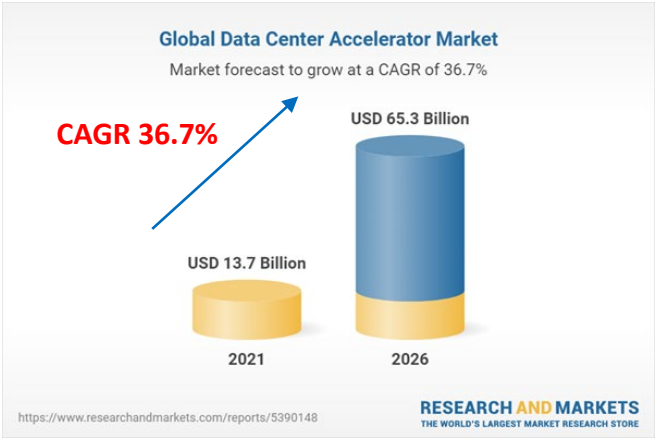
**Transforming Storage Rack to Data Service Stack**

# Data Centric Computing: Storage + Compute




***Market Size forecast 2026***
*Source: Global Market Insights/Research and Markets*

# Conclusion

- Data centric computing is very important for data intensive workload
  - Overcomes data gravity
- Moving computation closer to data is more efficient than transferring large amounts of data
- Federated data processing allows for better system efficiency
  - Reduced network traffic, less data movements, less time consumed
  - Lower latency
  - Improved total cost of ownership (TCO).
- Helps tackle the coexistence of ‘Big and Fast’ data challenges.

SNIA COMPUTE + MEMORY + STORAGE SUMMIT

*Architectures, Solutions, and Community*
*VIRTUAL EVENT, APRIL 11-12, 2023*

# Thank You!
# Please take a moment to rate this session.

Your feedback is important to us.