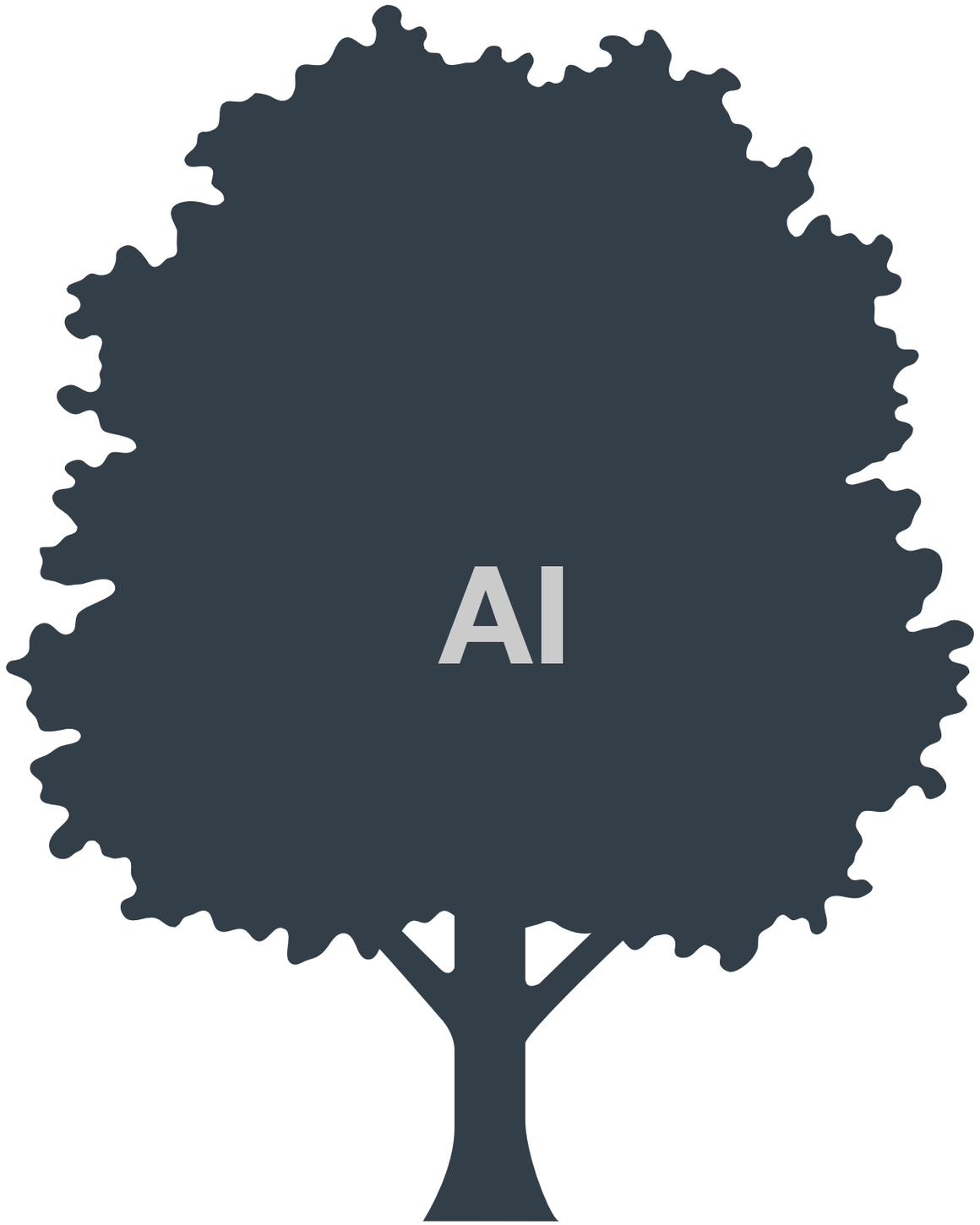
Missing the Forest for the Trees: End-to-End AI Application Performance in Edge Data Centers

International Symposium on High Performance Computer Architecture
25 February 2020

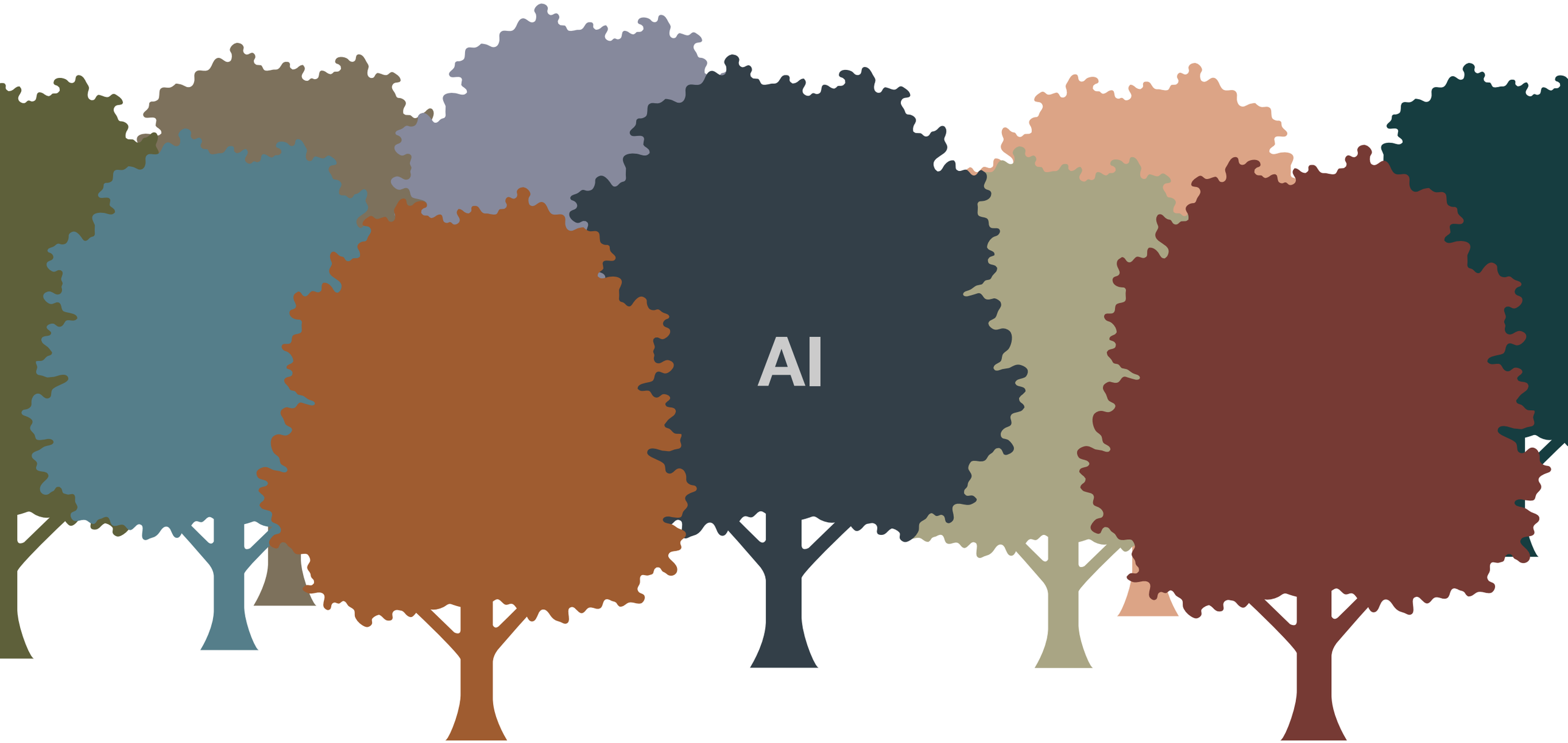
Daniel Richins[1,2], Dharmisha Doshi[2], Matthew Blackmore[2], Aswathy Thulaseedharan Nair[2], Neha Pathapati[2], Ankit Patel[2], Brainard Daguman[2], Daniel Dobrijalowski[2], Ramesh Illikkal[2], Kevin Long[2], David Zimmerman[2], Vijay Janapa Reddi[1,3]

[1]The University of Texas at Austin
[2]Intel
[3]Harvard University

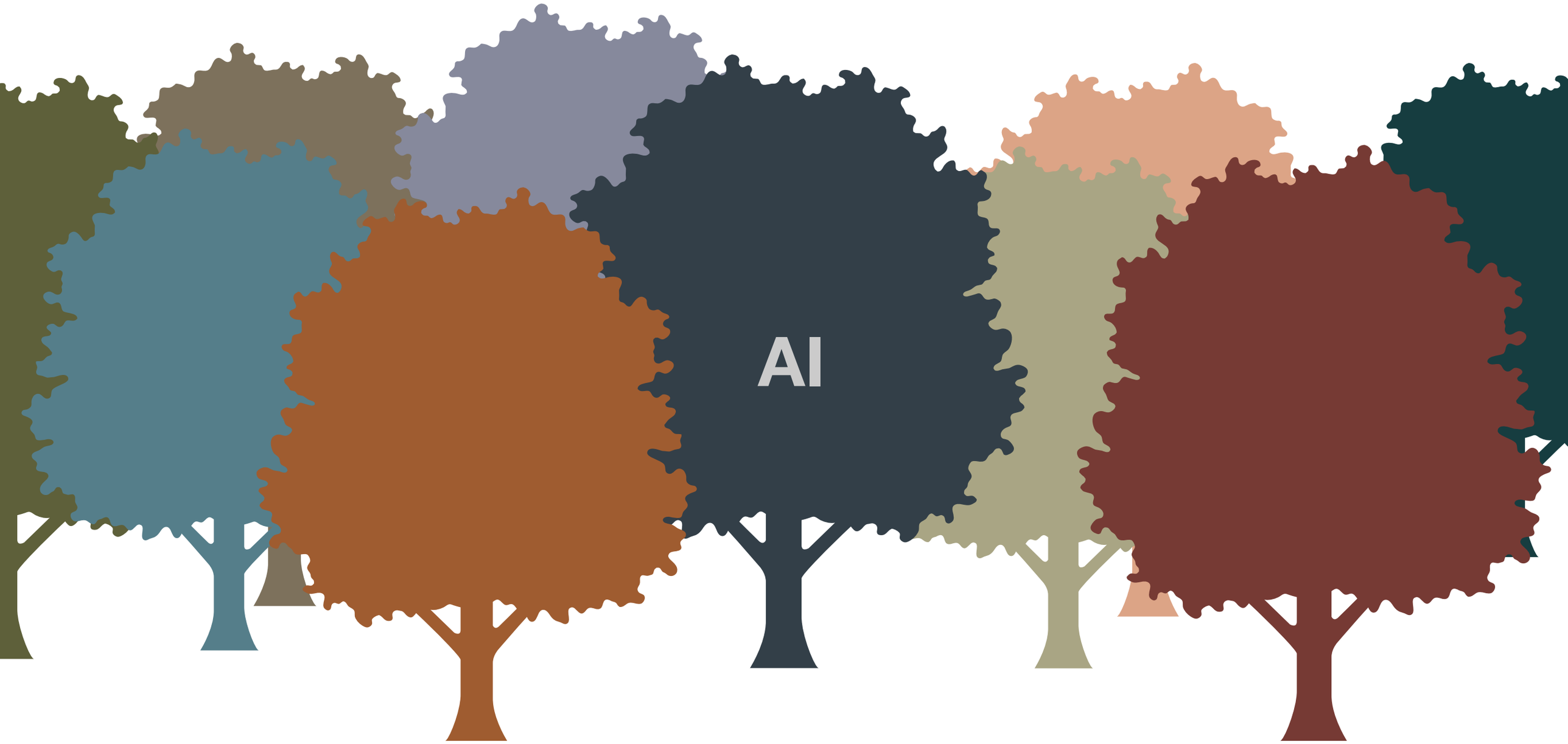
Missing the Forest for the Trees
The AI Tax



Missing the Forest for the Trees
The AI Tax



Missing the Forest for the Trees
The AI Tax



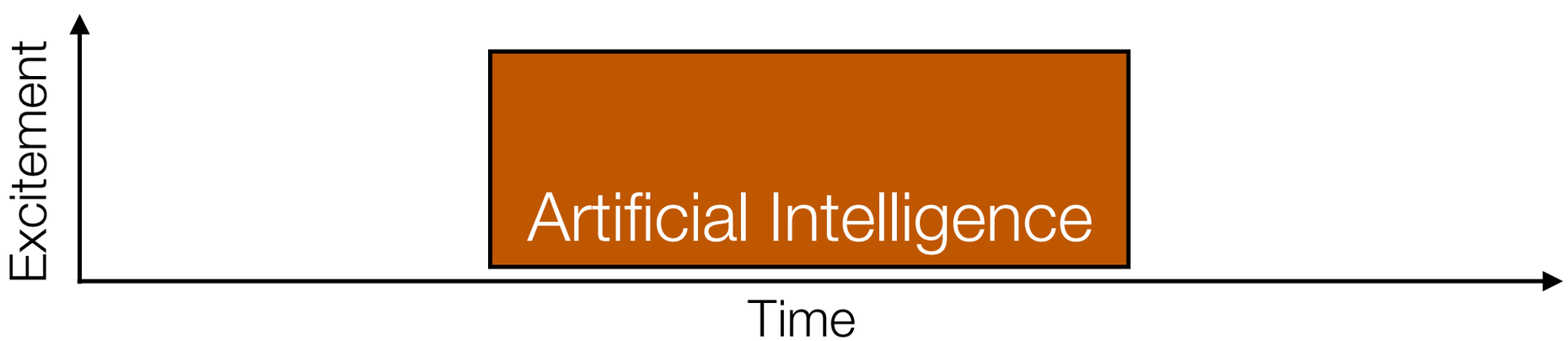
The forest is the **AI tax**

Missing the Forest for the Trees
The AI Tax

The AI tax includes all the compute and infrastructure in an AI application that is necessary to *enable* the AI to execute but that isn't AI itself.

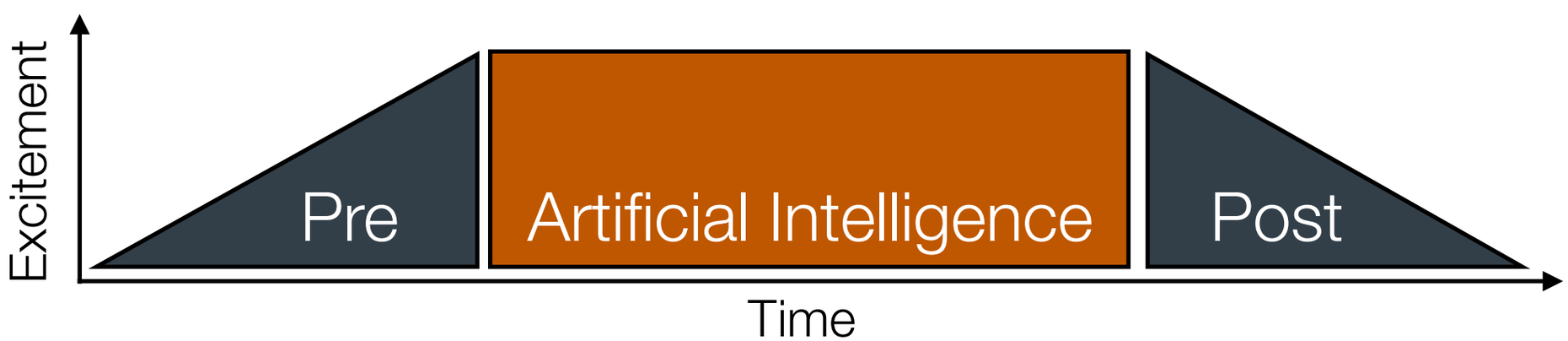
Missing the Forest for the Trees
The AI Tax

The AI tax includes all the compute and infrastructure in an AI application that is necessary to *enable* the AI to execute but that isn't AI itself.



Missing the Forest for the Trees
The AI Tax

The AI tax includes all the compute and infrastructure in an AI application that is necessary to *enable* the AI to execute but that isn't AI itself.



1. AI Tax
    a. Definition
    b. Video Analytics
    c. Analysis
2. AI Acceleration
    a. Emulation Technique
    b. Results
    c. What's Breaking?
3. Optimization
    a. Fixing the Bottleneck
    b. Edge Data Centers
    c. Two Designs
4. Conclusion

Outline

1. **AI Tax - A Case Study**
    a. Definition
    b. Video Analytics
    c. Analysis
2. **AI Acceleration**
    a. Emulation Technique
    b. Results
    c. What's Breaking?
3. **Optimization**
    a. Fixing the Bottleneck
    b. Edge Data Centers
    c. Two Designs
4. **Conclusion**

Outline

1. **AI Tax - A Case Study**
    a. **Definition**
    b. **Video Analytics**
    c. **Analysis**
2. **AI Acceleration - Anticipating Future Bottlenecks**
    a. **Emulation Technique**
    b. **Results**
    c. **What's Breaking?**
3. **Optimization**
    a. **Fixing the Bottleneck**
    b. **Edge Data Centers**
    c. **Two Designs**
4. **Conclusion**

Outline

1. AI Tax - A Case Study
    a. Definition
    b. Video Analytics
    c. Analysis
2. AI Acceleration - Anticipating Future Bottlenecks
    a. Emulation Technique
    b. Results
    c. What's Breaking?
3. Optimization - Better Performance at Lower TCO
    a. Fixing the Bottleneck
    b. Edge Data Centers
    c. Two Designs
4. Conclusion

Outline

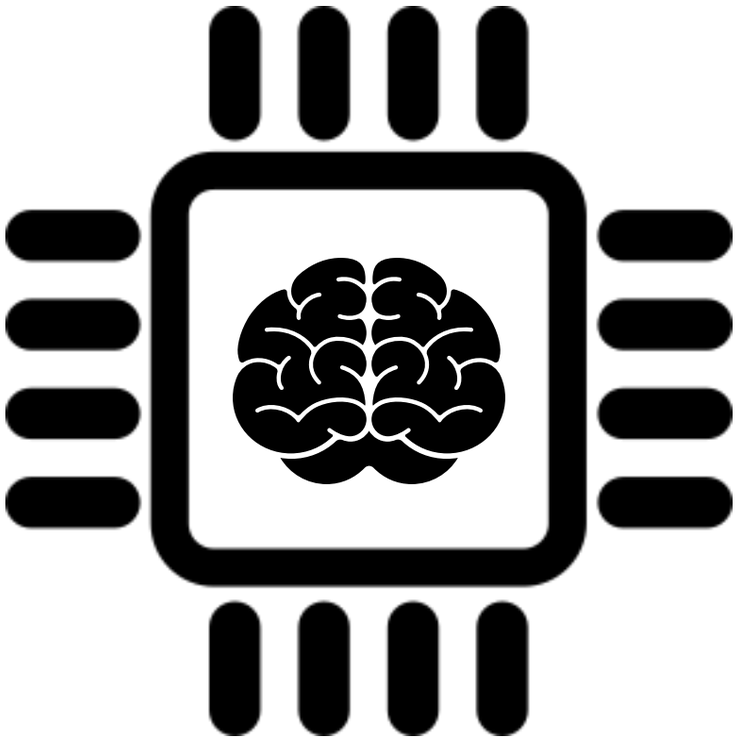
1. AI Tax
    a. Definition
    b. Video Analytics
    c. Analysis
2. AI Acceleration
    a. Emulation Technique
    b. Results
    c. What's Breaking?
3. Optimization
    a. Fixing the Bottleneck
    b. Edge Data Centers
    c. Two Designs
4. Conclusion

AI Tax

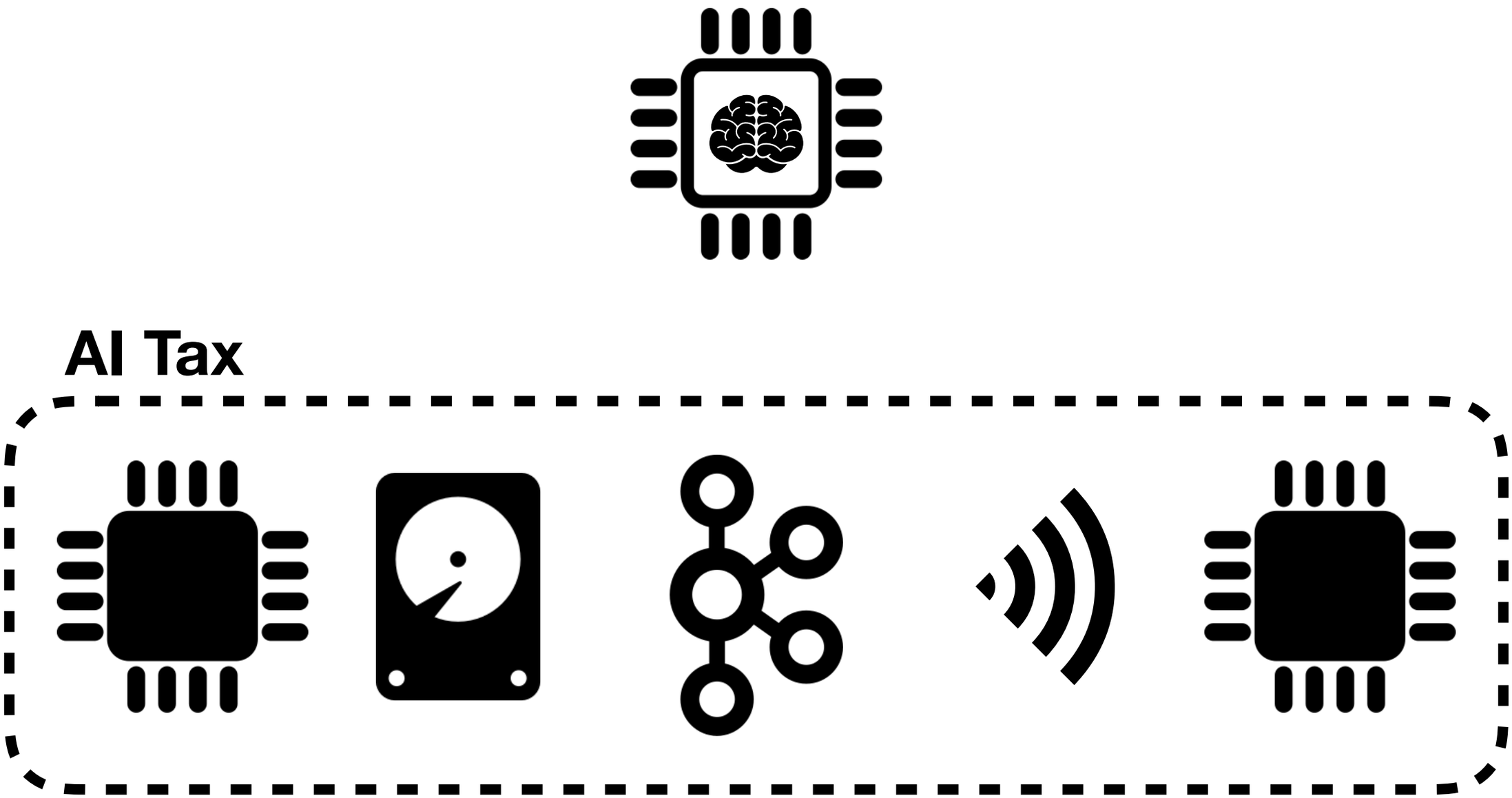
AI Tax - Definition

Missing the Forest for the Trees

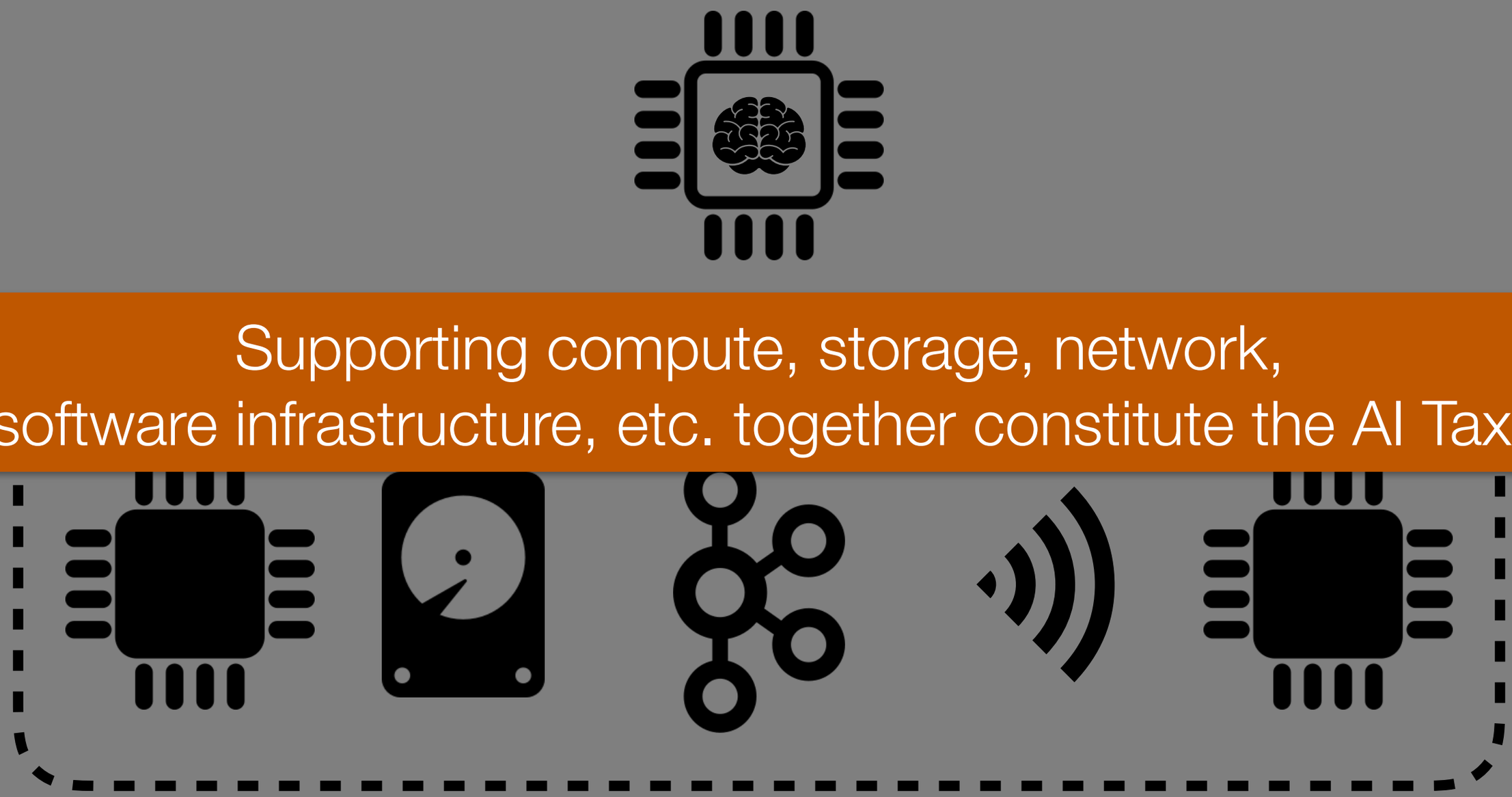
Missing the Forest for the Trees



Missing the Forest for the Trees



Missing the Forest for the Trees



Supporting compute, storage, network, software infrastructure, etc. together constitute the AI Tax.

AI Tax - Video Analytics

Face Recognition Algorithm

Face Recognition Algorithm

Face Recognition is Google's FaceNet as a data center application.

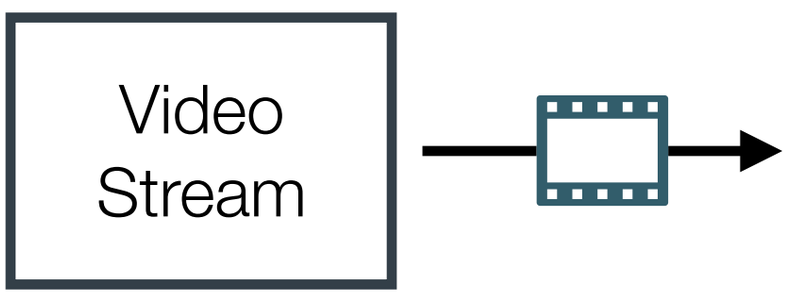
Face Recognition Algorithm

Face Recognition is Google's FaceNet as a data center application.

Video Stream

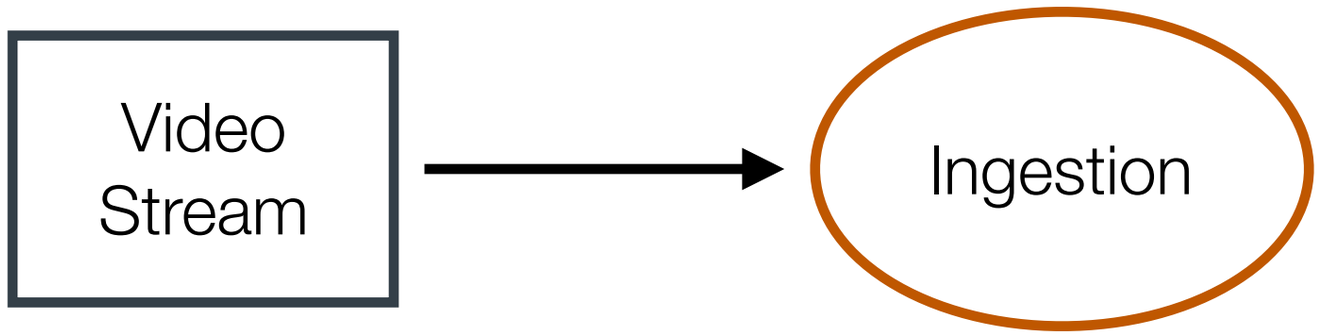
Face Recognition Algorithm

Face Recognition is Google's FaceNet as a data center application.



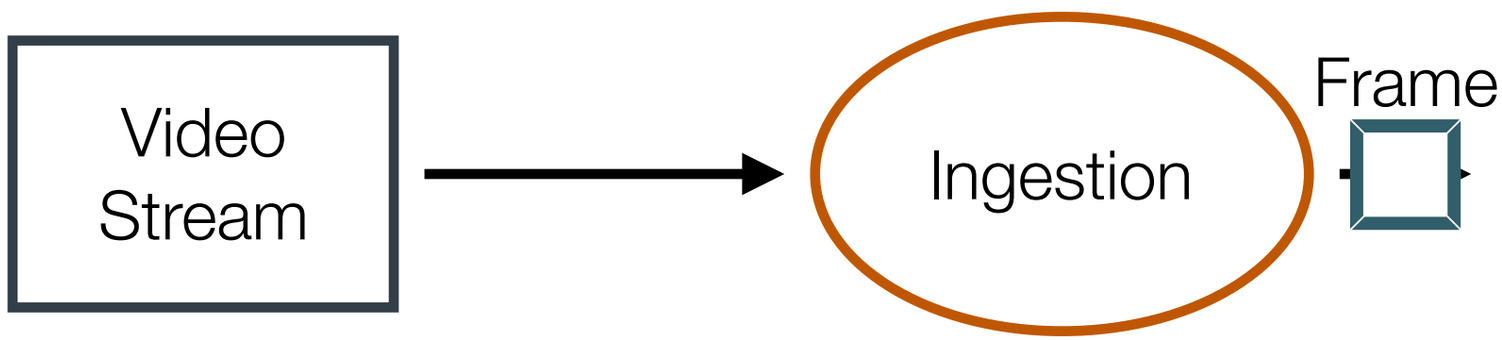
Face Recognition Algorithm

Face Recognition is Google's FaceNet as a data center application.



Face Recognition Algorithm

Face Recognition is Google's FaceNet as a data center application.

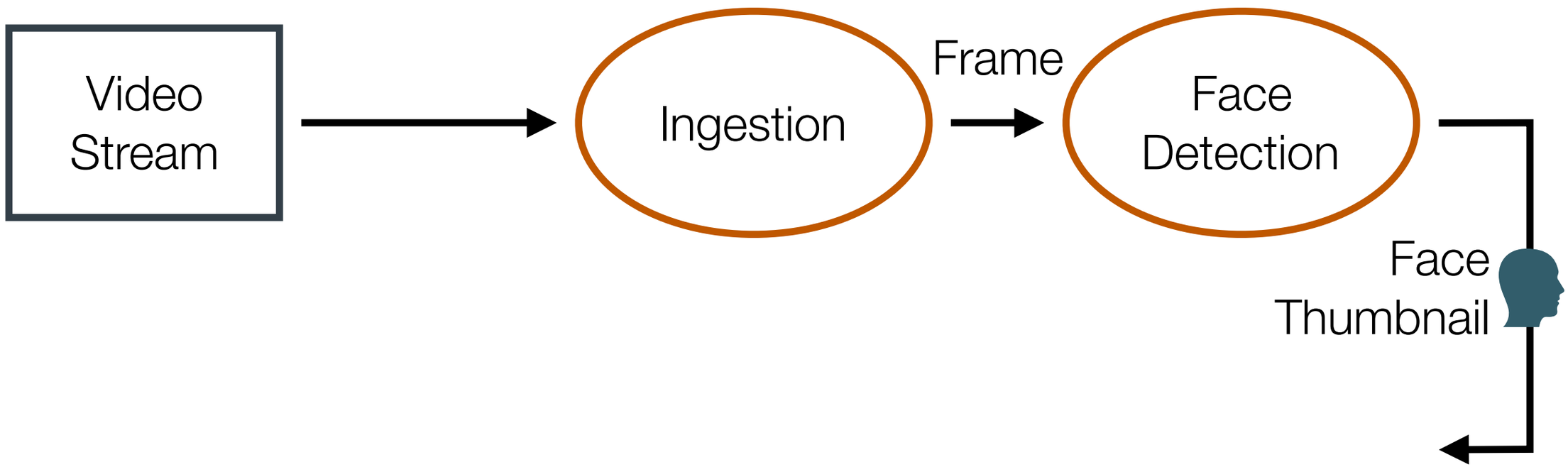


Face Recognition Algorithm

Face Recognition is Google's FaceNet as a data center application.

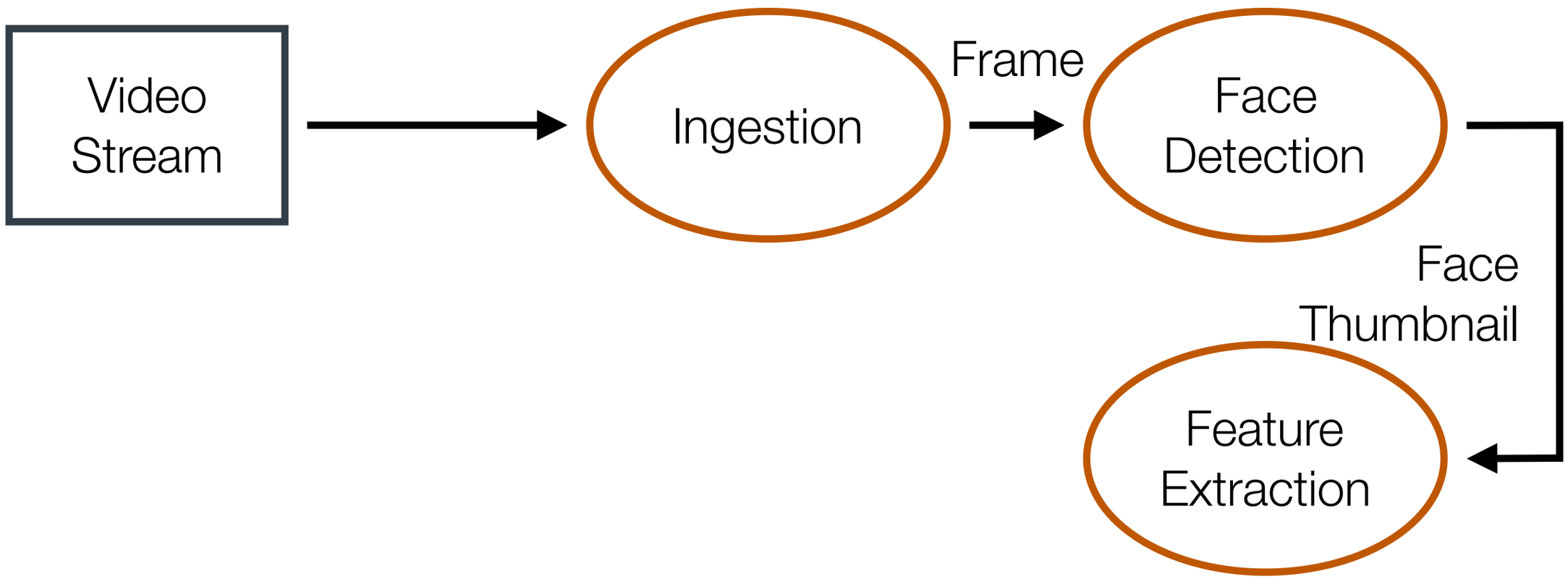
Face Recognition Algorithm

Face Recognition is Google's FaceNet as a data center application.



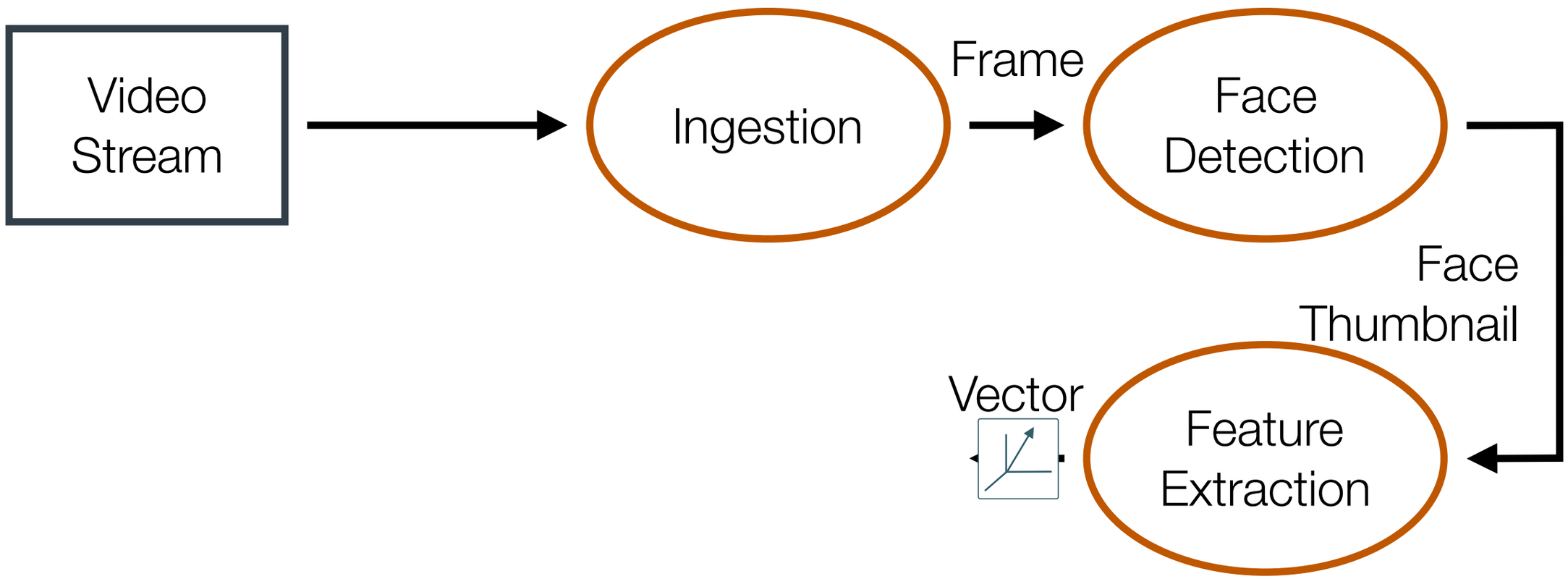
Face Recognition Algorithm

Face Recognition is Google's FaceNet as a data center application.



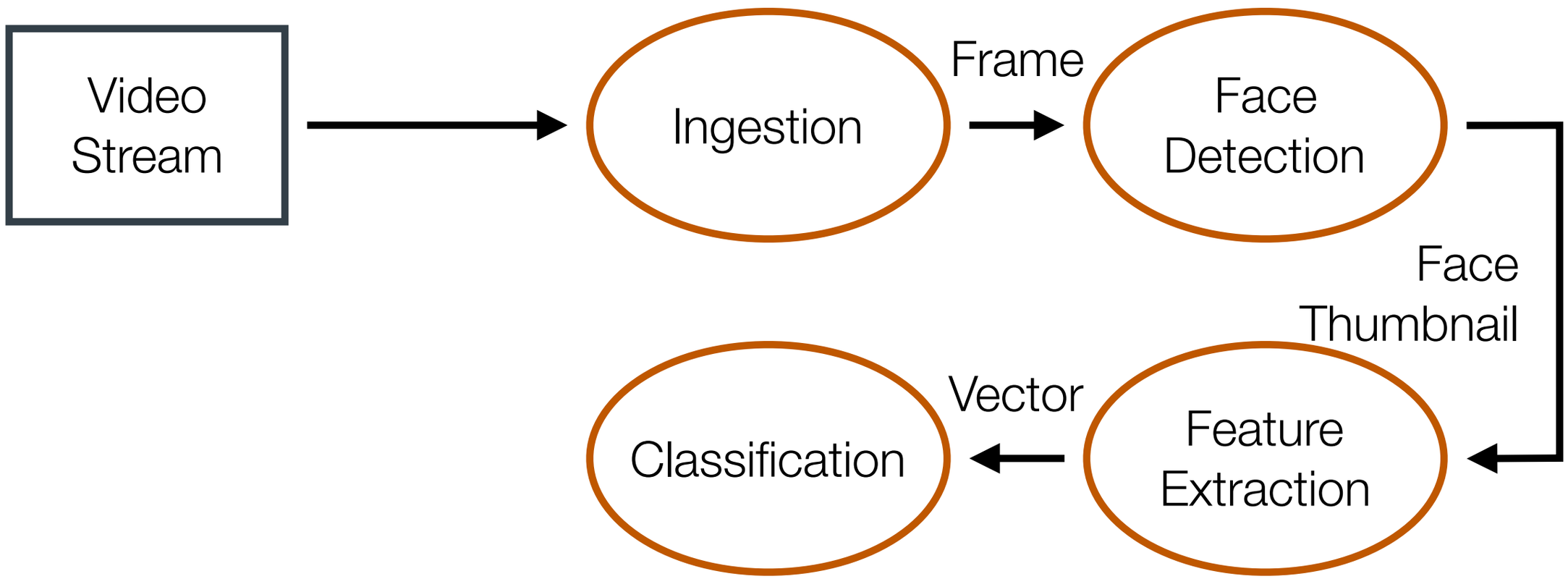
Face Recognition Algorithm

Face Recognition is Google's FaceNet as a data center application.



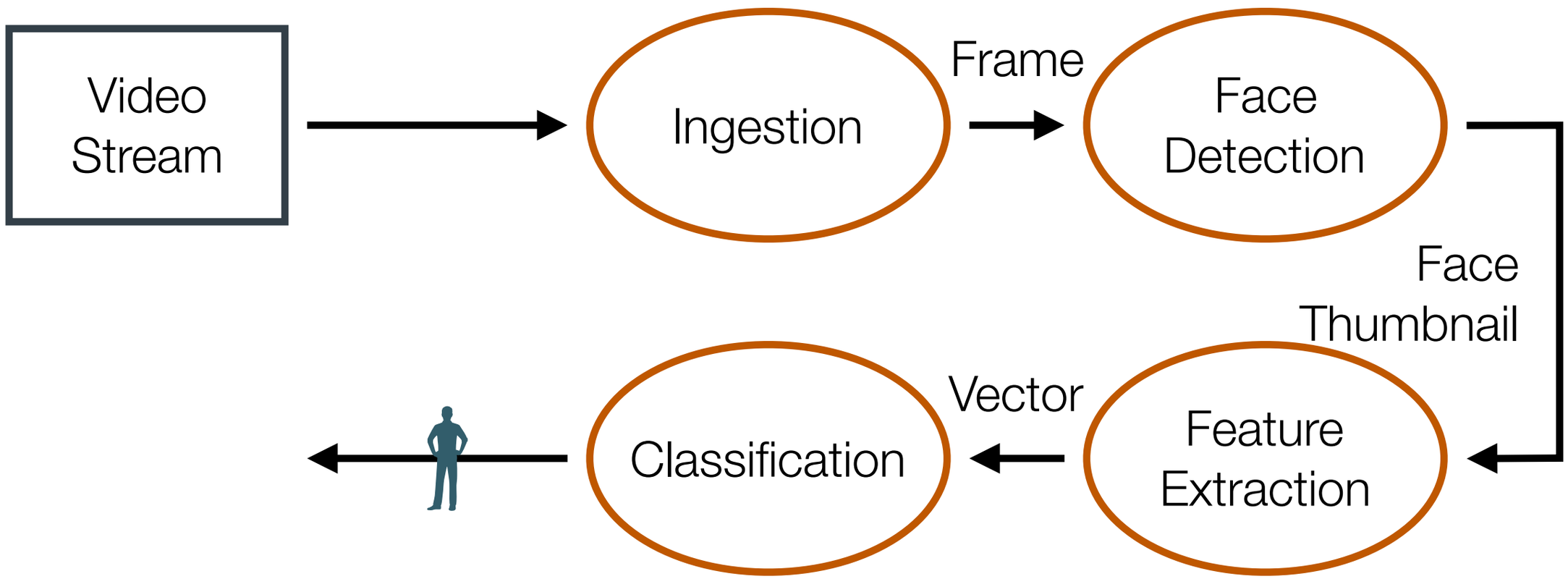
Face Recognition Algorithm

Face Recognition is Google's FaceNet as a data center application.



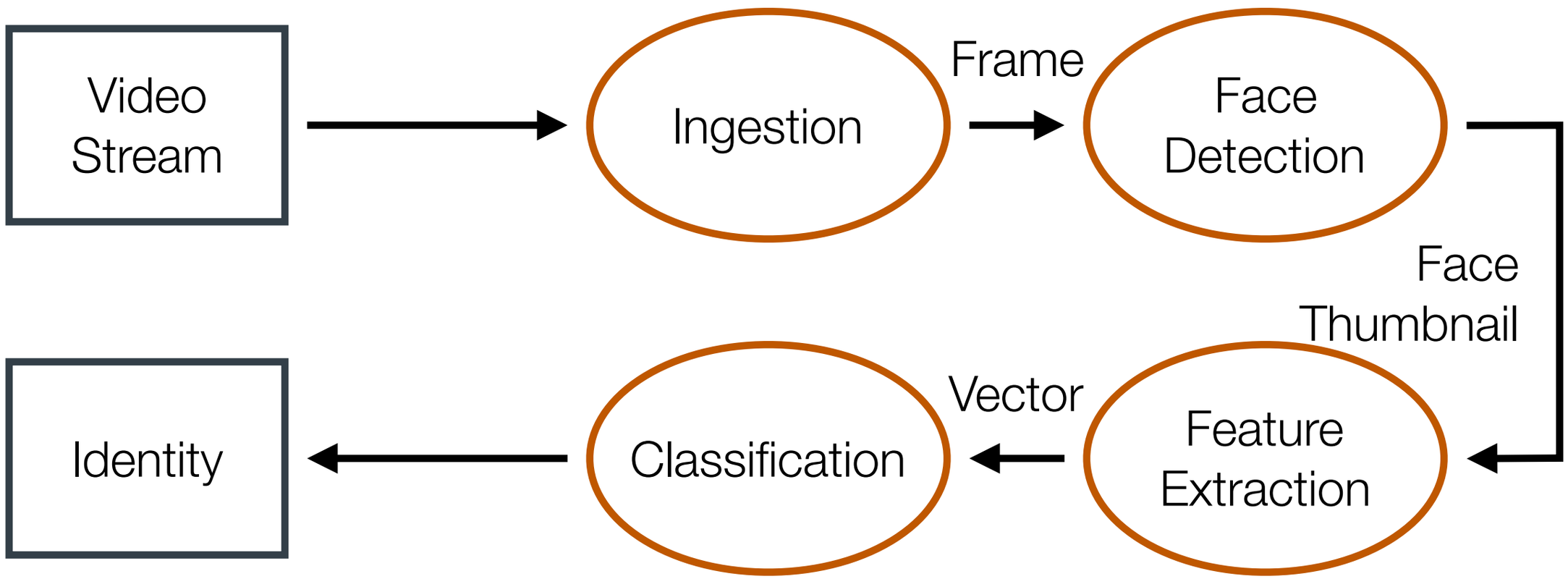
Face Recognition Algorithm

Face Recognition is Google's FaceNet as a data center application.



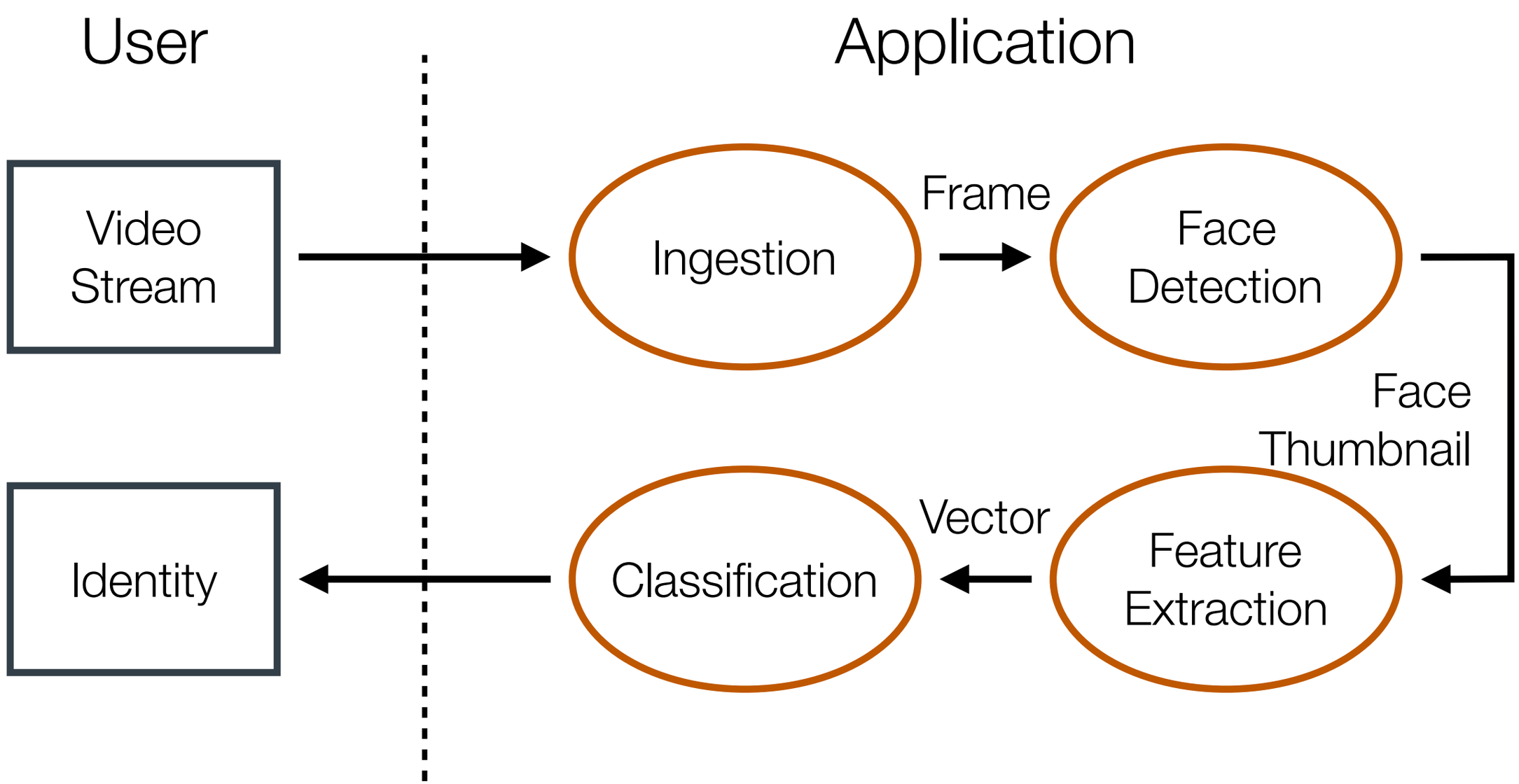
Face Recognition Algorithm

Face Recognition is Google's FaceNet as a data center application.



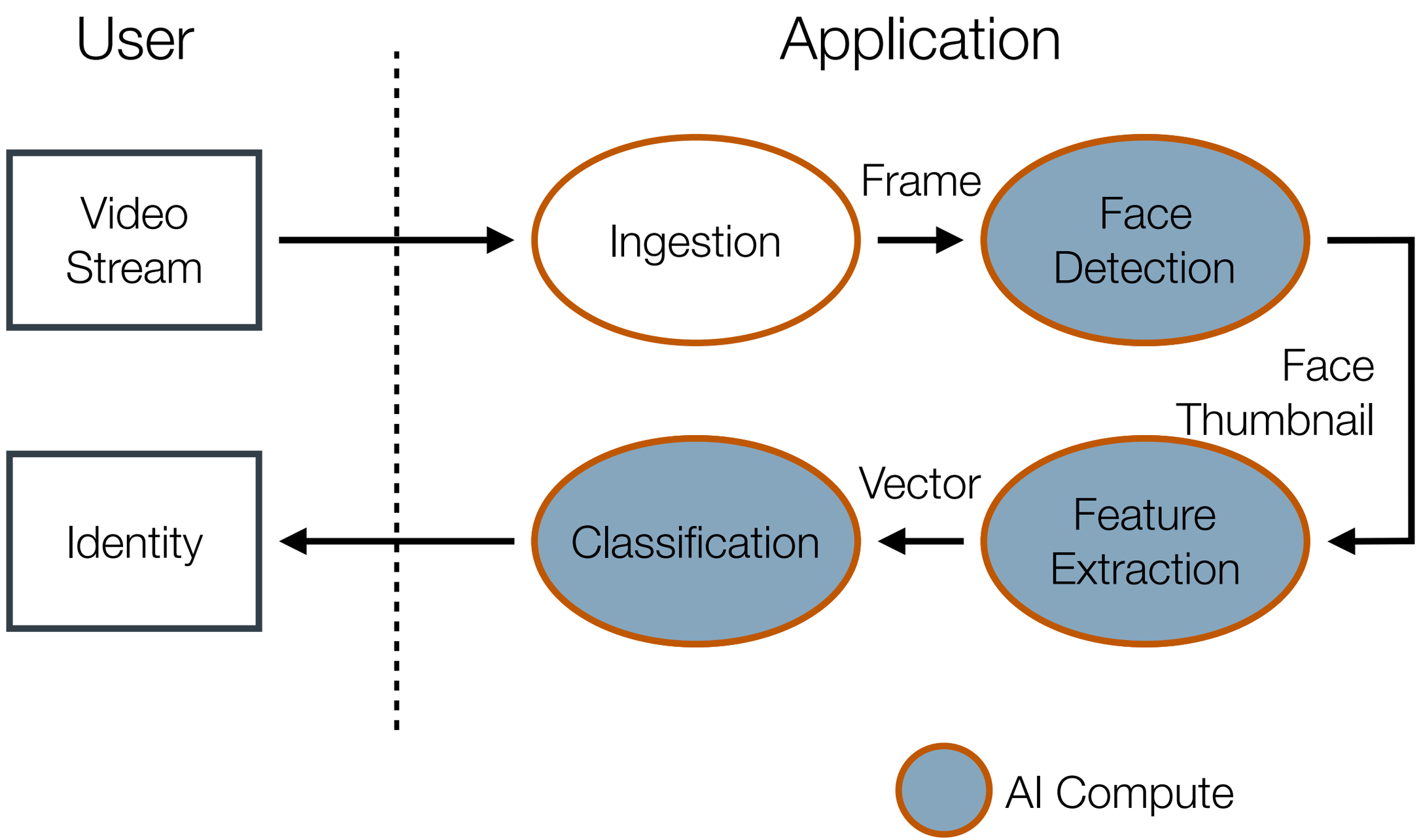
Face Recognition Algorithm

Face Recognition is Google's FaceNet as a data center application.

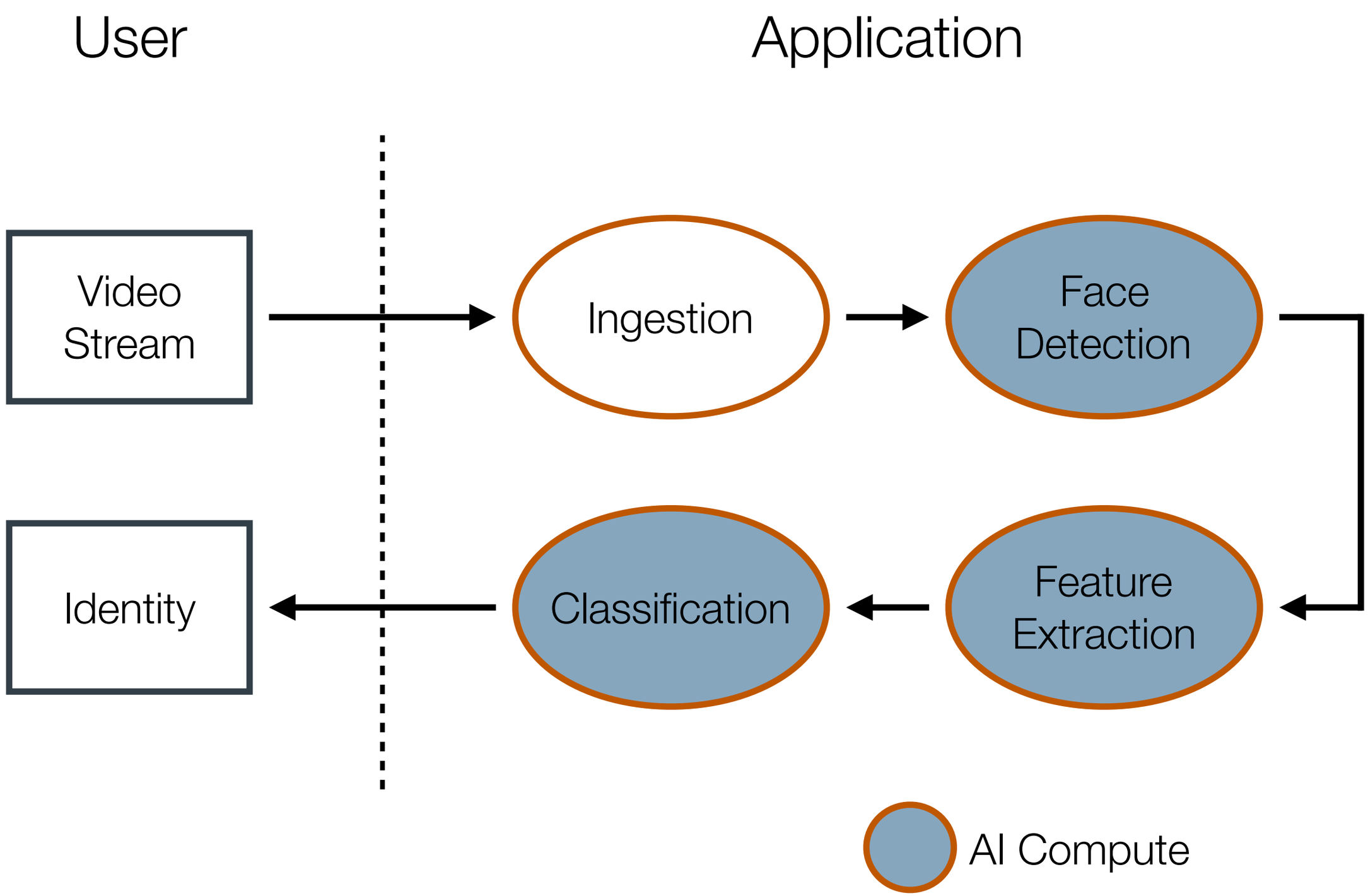


Face Recognition Algorithm

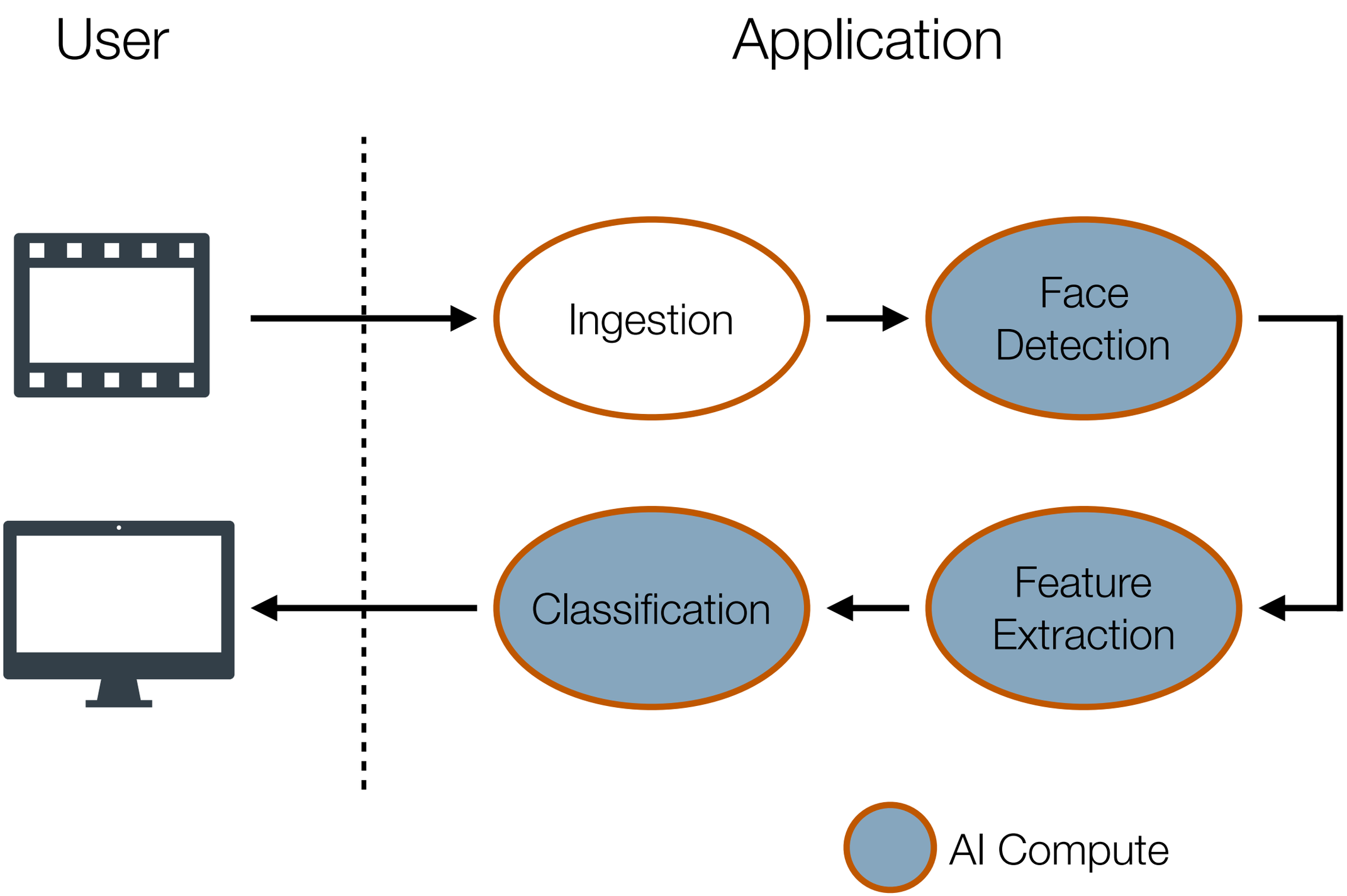
Face Recognition is Google's FaceNet as a data center application.



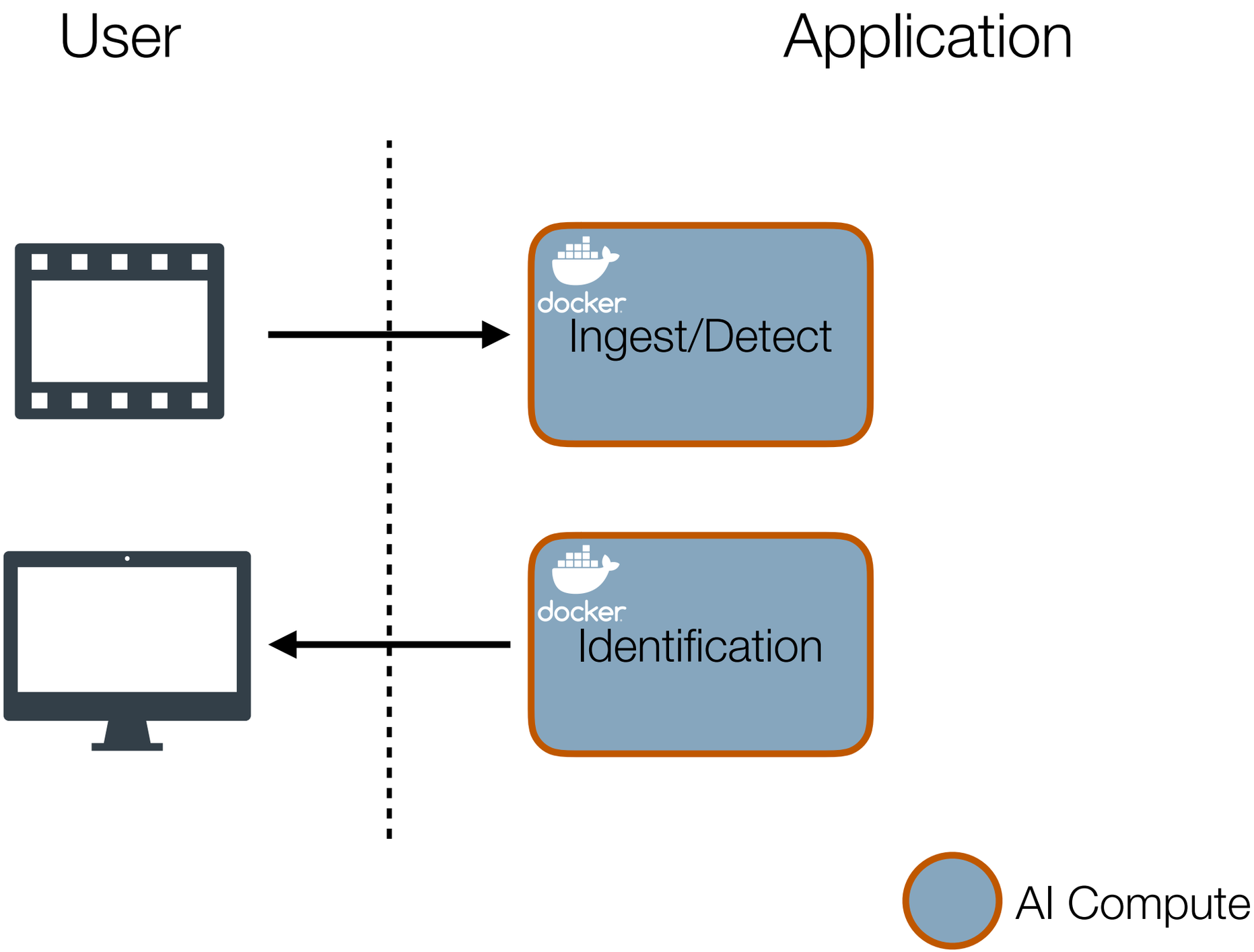
Face Recognition
Data Center Deployment



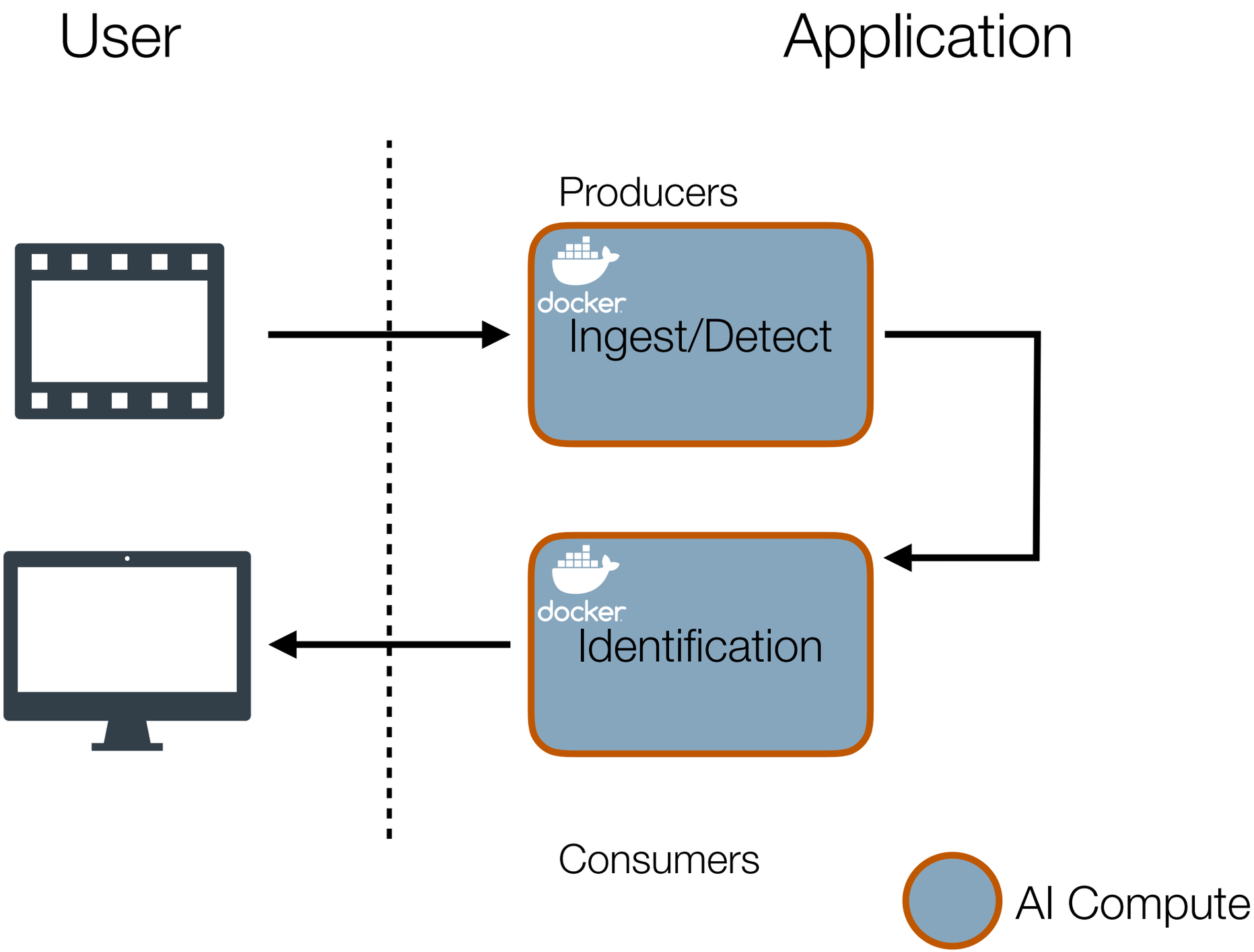
Face Recognition
Data Center Deployment



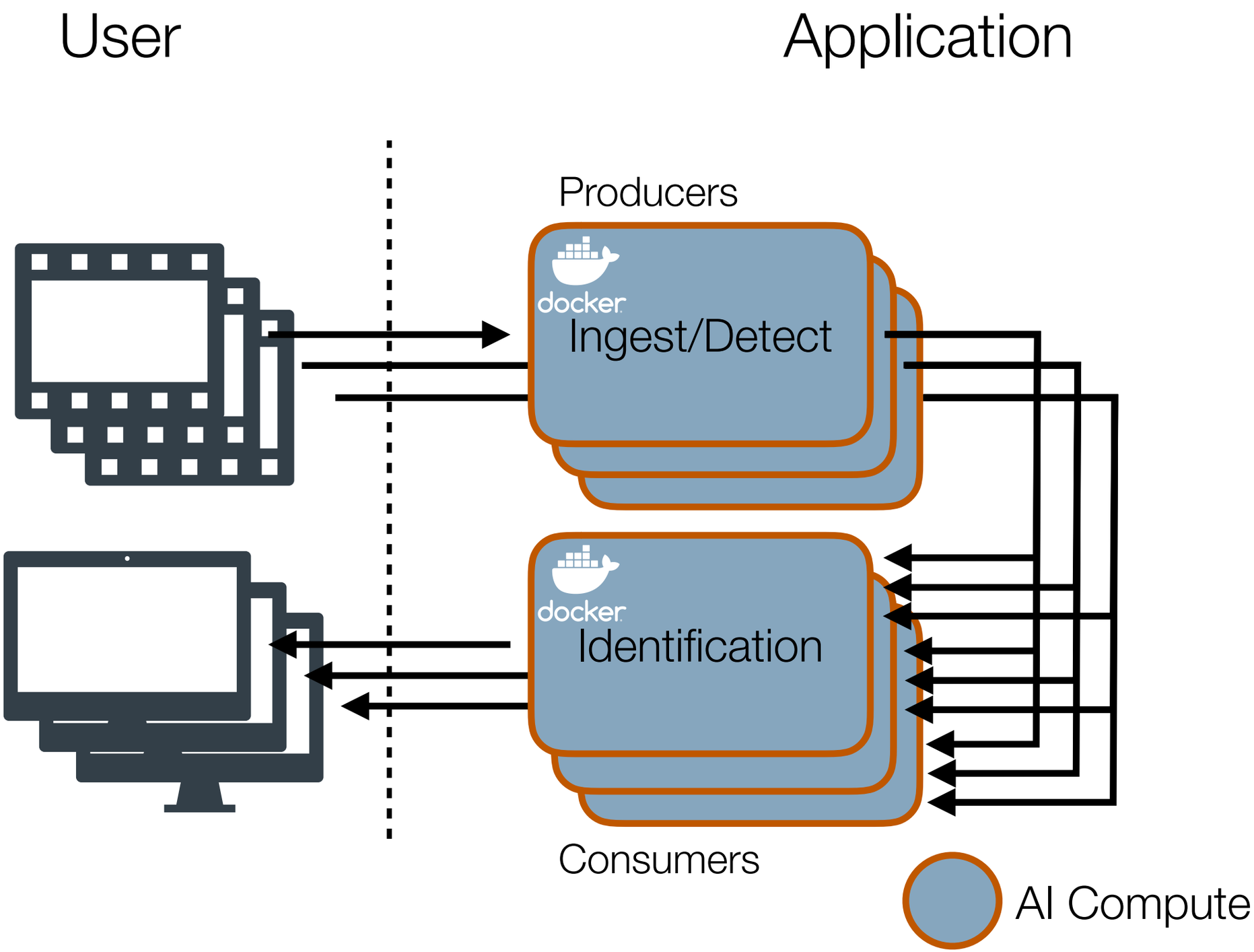
Face Recognition
Data Center Deployment



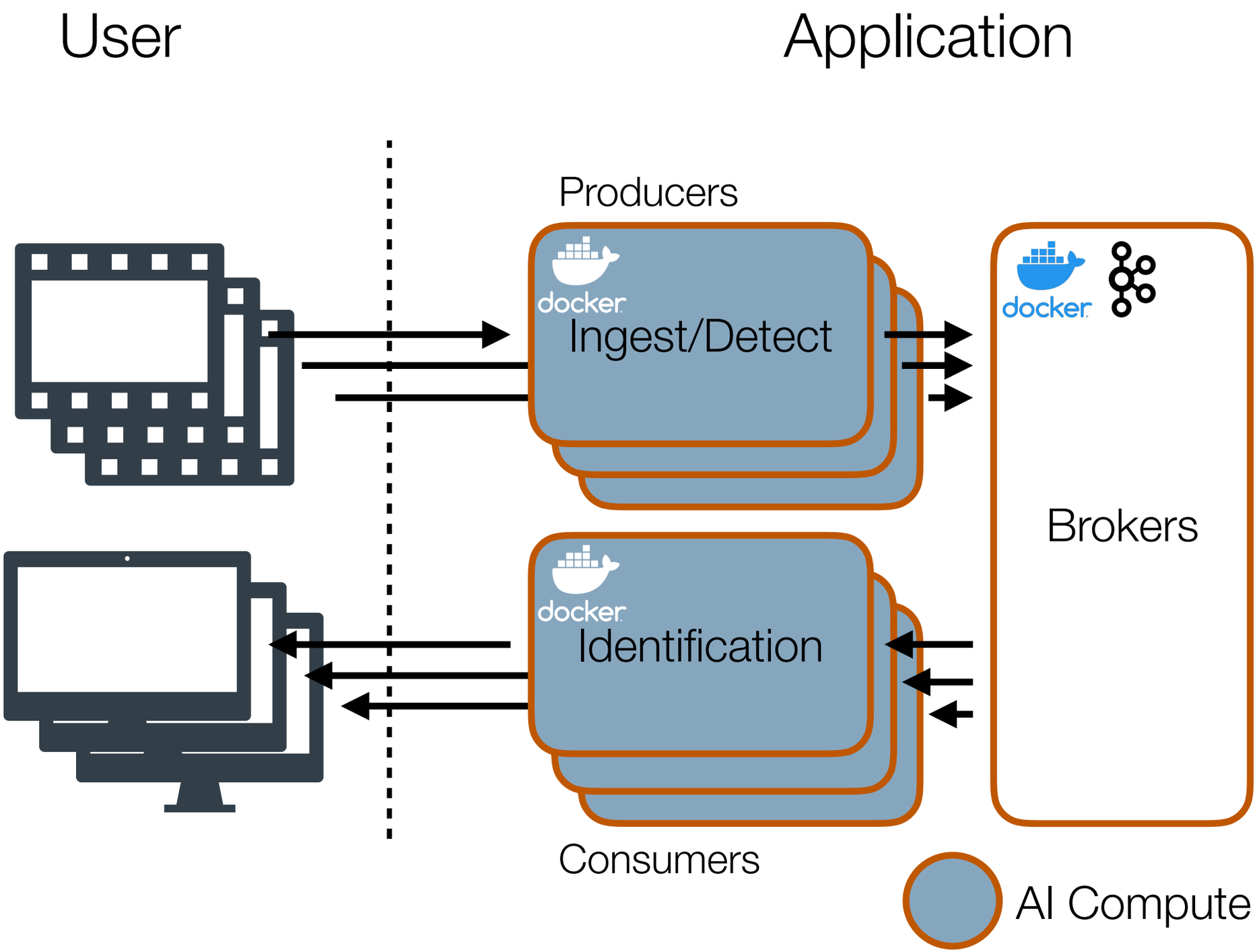
Face Recognition
Data Center Deployment



Face Recognition
Data Center Deployment

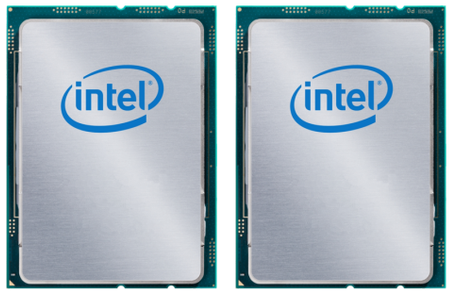


Face Recognition
Data Center Deployment

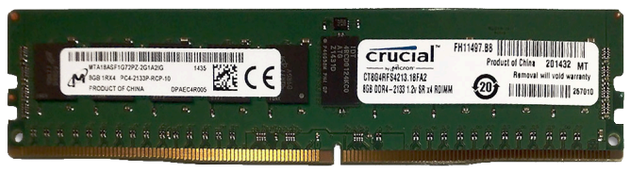


Experimental Setup
Hardware

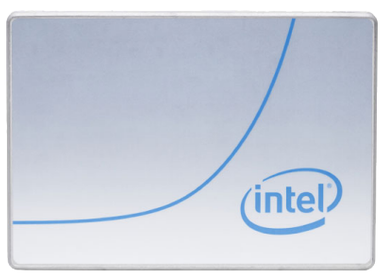
Experimental Setup
Hardware



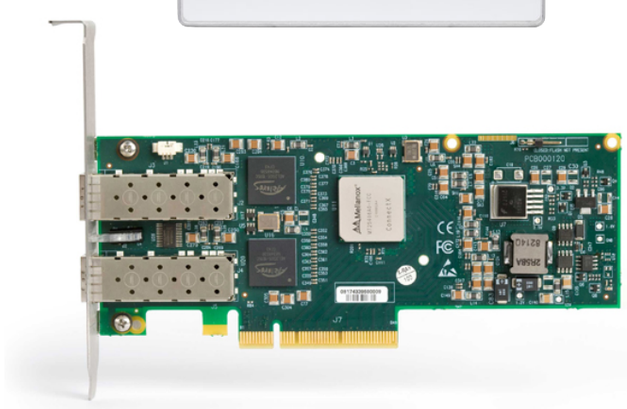
2x Intel Xeon Platinum 8176
2x 28 cores, 2.10 GHz, 2x 38.5 MB LLC



384 GB DDR4 SDRAM



1x Intel SSD P4510
2.85 GB/s read
1.10 GB/s write



100 Gbps Ethernet

Experimental Setup
Hardware

Experimental Setup
Face Recognition

We allocate one core per container.

Hence, a server runs 56 containers.

Experimental Setup
Face Recognition

We allocate one core per container.

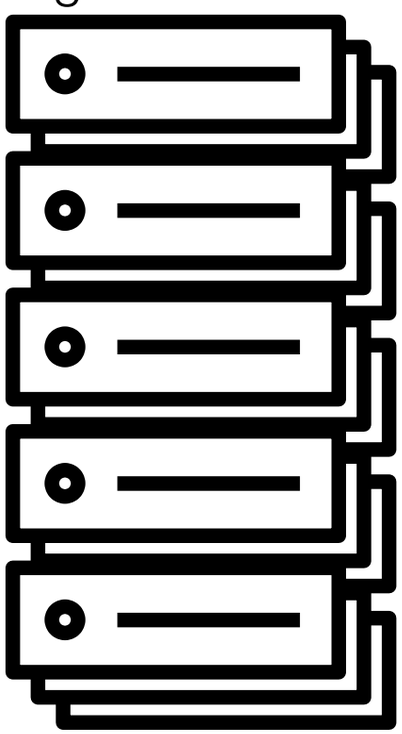
Hence, a server runs 56 containers.

Experimental Setup
Face Recognition

We allocate one core per container.

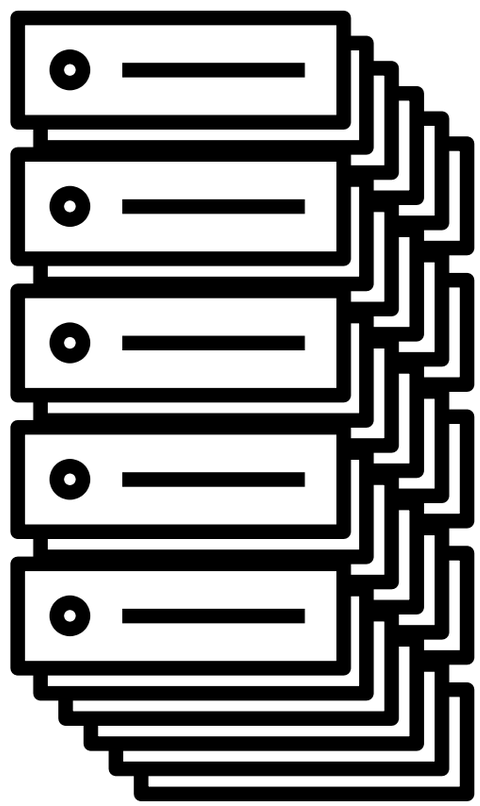
Hence, a server runs 56 containers.

Ingest/Detect



840 total producers

Identification



1680 total consumers

Experimental Setup
Face Recognition

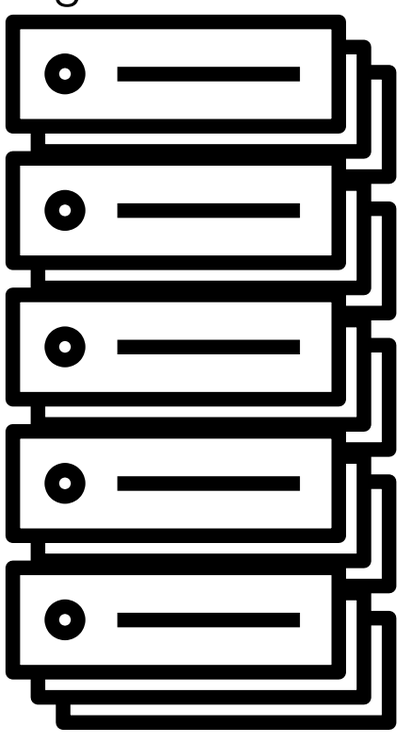
We allocate one core per container.

Hence, a server runs 56 containers.

Brokers get their own server.

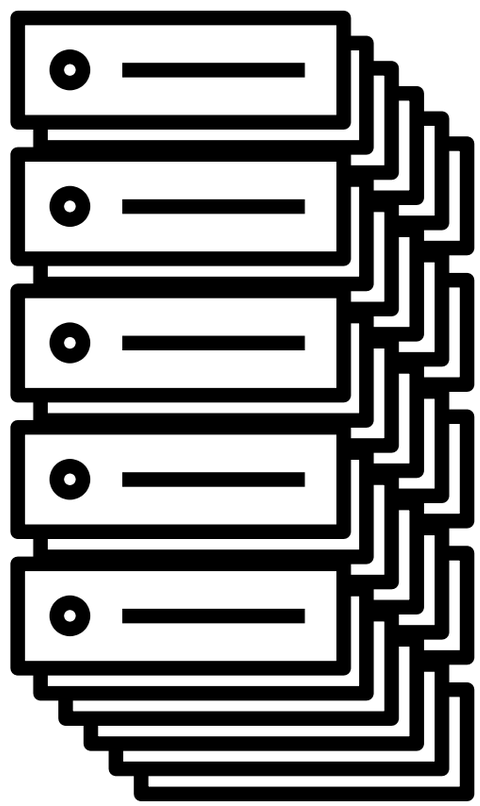
This grants them full network and storage bandwidth.

Ingest/Detect



840 total producers

Identification



1680 total consumers

Experimental Setup
Face Recognition

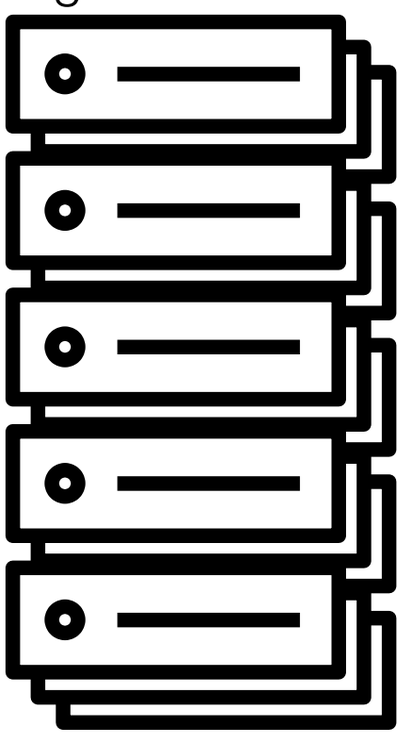
We allocate one core per container.

Hence, a server runs 56 containers.

Brokers get their own server.

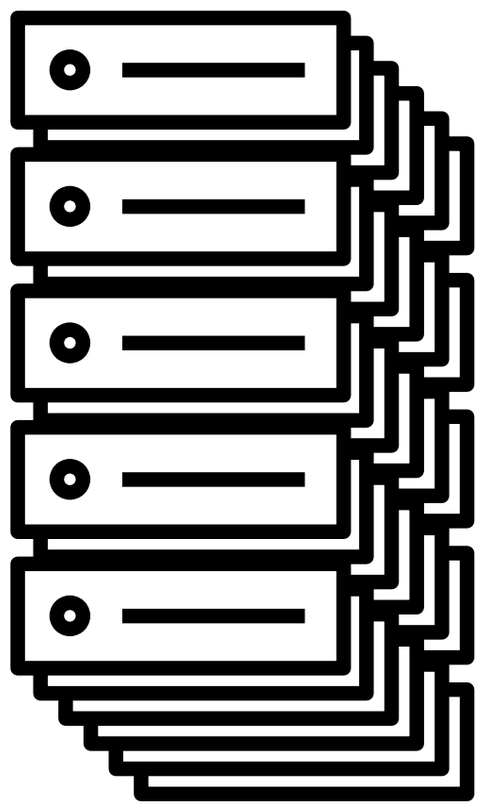
This grants them full network and storage bandwidth.

Ingest/Detect

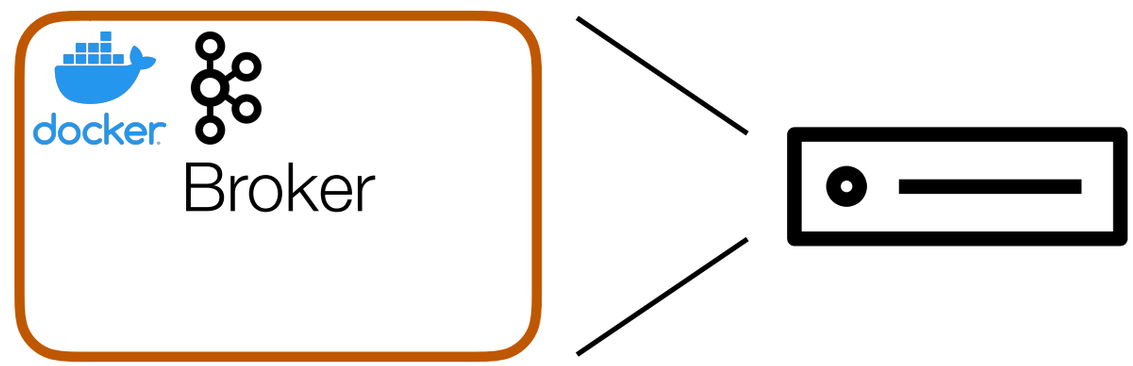


840 total producers

Identification



1680 total consumers



Experimental Setup
Face Recognition

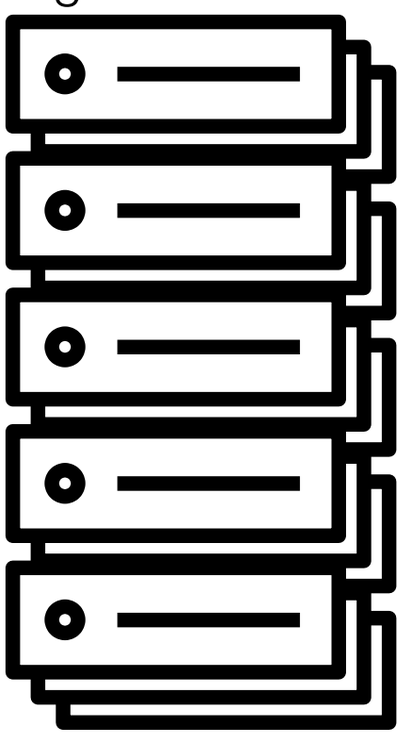
We allocate one core per container.

Hence, a server runs 56 containers.

Brokers get their own server.

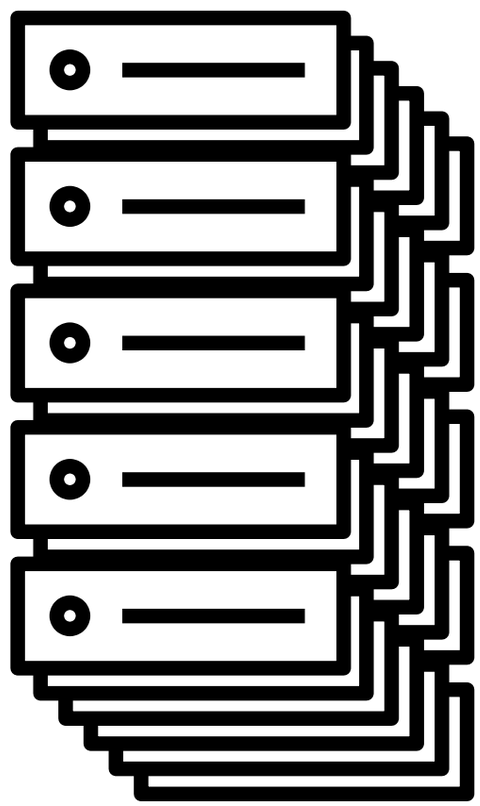
This grants them full network and storage bandwidth.

Ingest/Detect



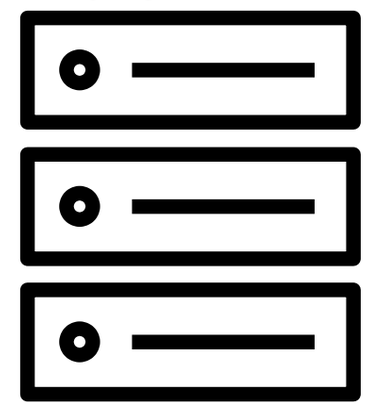
840 total producers

Identification



1680 total consumers

Broker



3 brokers

AI Tax - Analysis

Application Progress
Event Logging

```
while True:
    frame = queue.get()
    faces = detect_faces(frame)
    producer.send(faces)
```

Application Progress Event Logging

```
while True:
    frame = queue.get()
    start = time.time()
    faces = detect_faces(frame)
    end = time.time()
    producer.send(faces)
```

Application Progress Event Logging

```
while True:
    frame = queue.get()
    start = time.time()
    faces = detect_faces(frame)
    end = time.time()
    producer.send(faces)
    size = sys.getsizeof(faces)
    log = {
        'start': start,
        'end': end,
        'size': size
    }
    logger.info(log)
```

Application Progress Event Logging

```
while True:
    frame = queue.get()
    start = time.time()
    faces = detect_faces(frame)
```

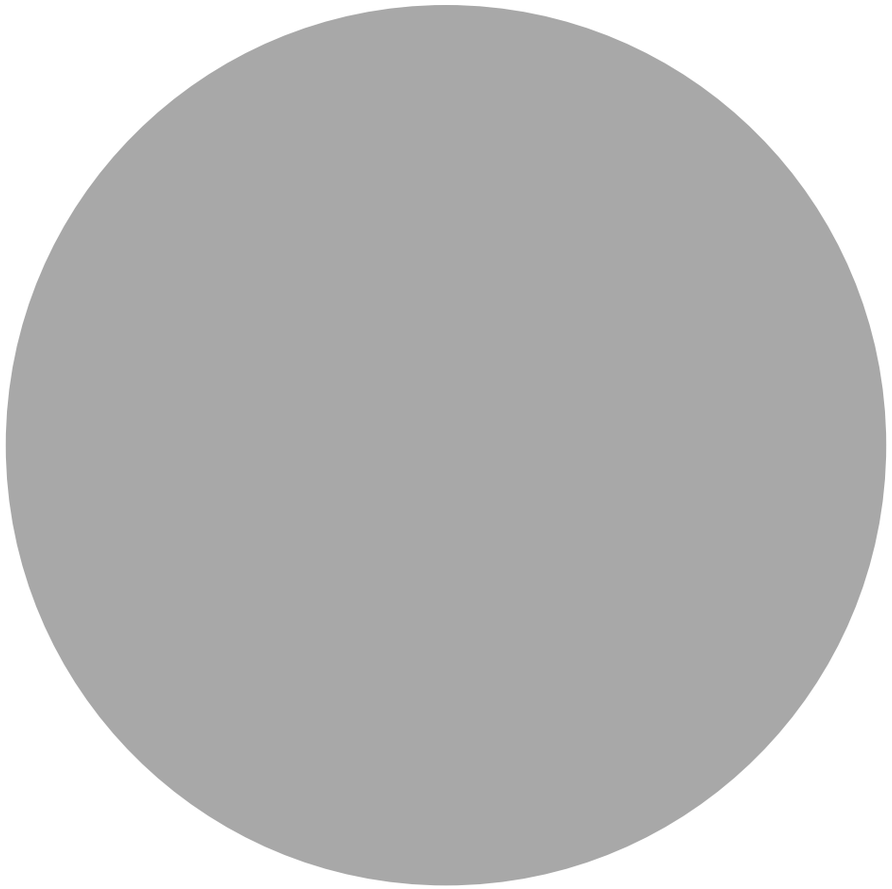
Logging is designed to raise the level of abstraction.
We view application progress from the data center perspective.

```
        'start': start,
        'end': end,
        'size': size
    }
    logger.info(log)
```

Face Detection Latency

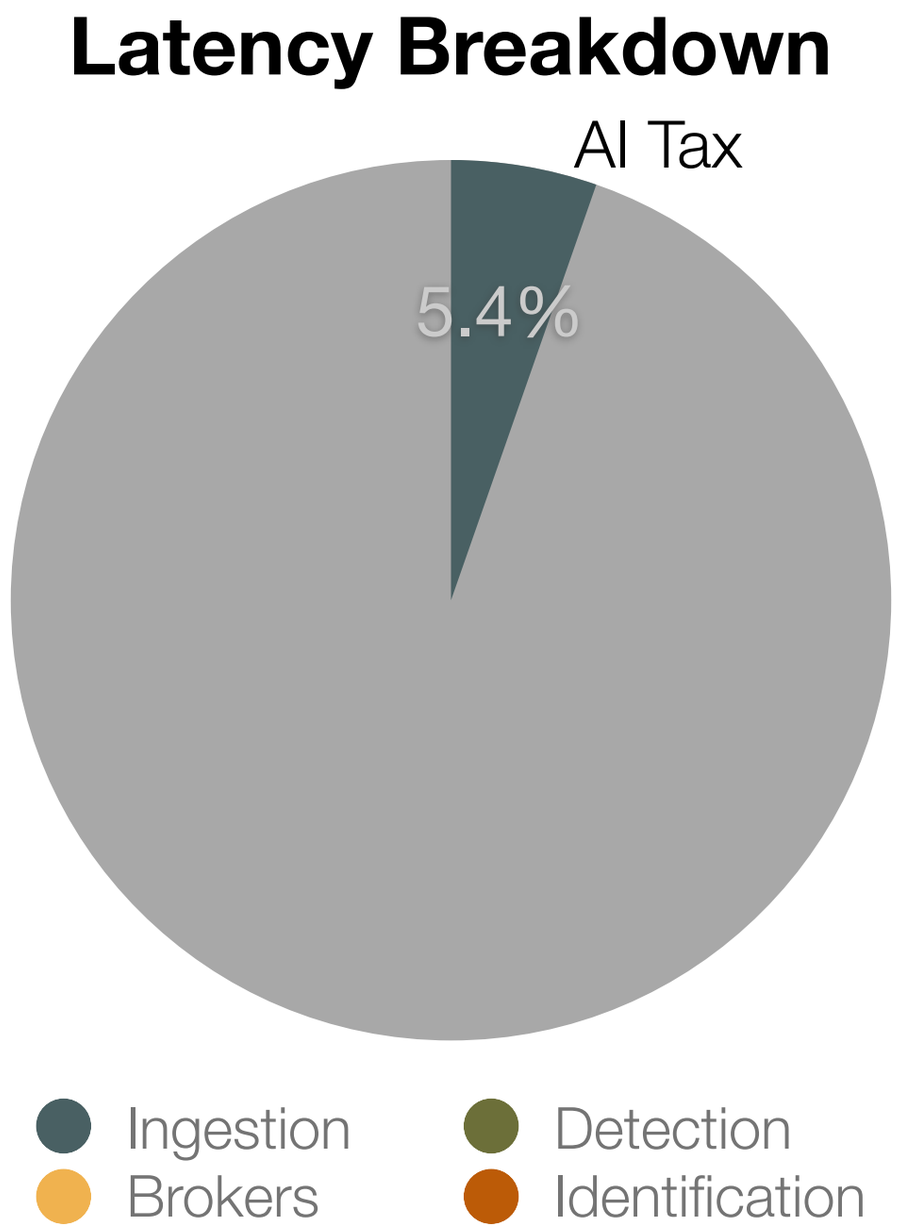
Face Detection Latency

Latency Breakdown

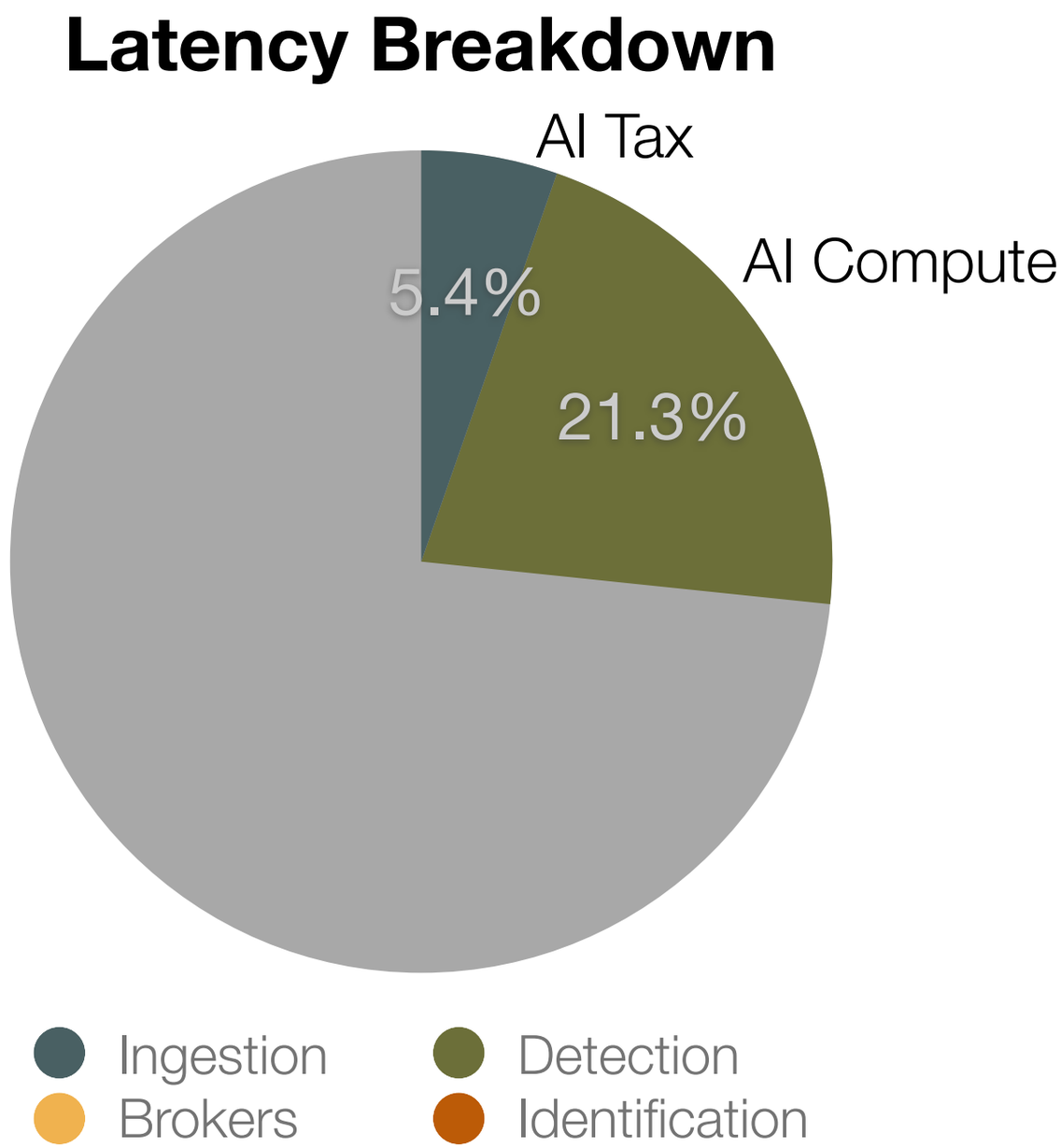


Ingestion
Brokers
Detection
Identification

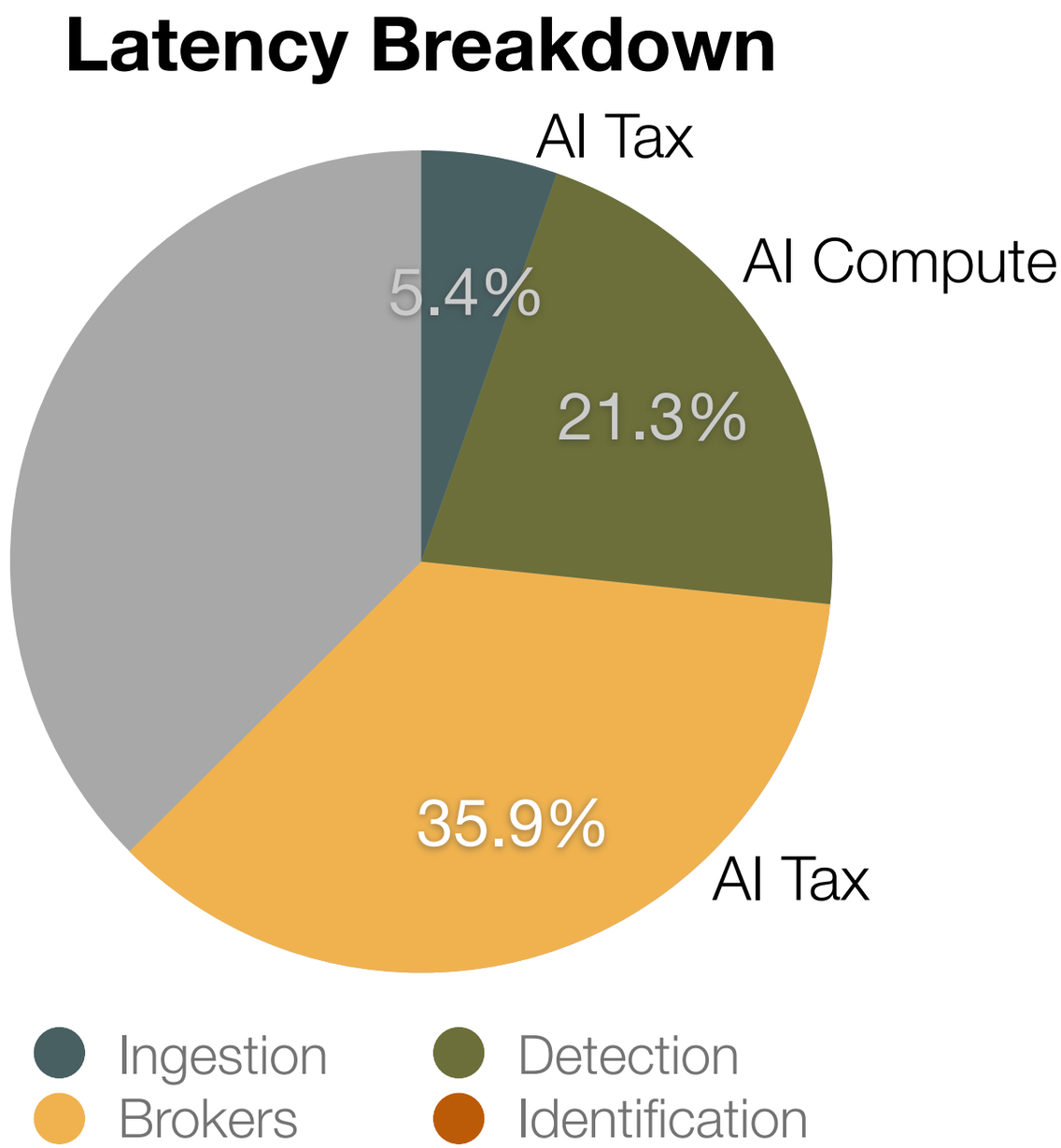
Face Detection Latency



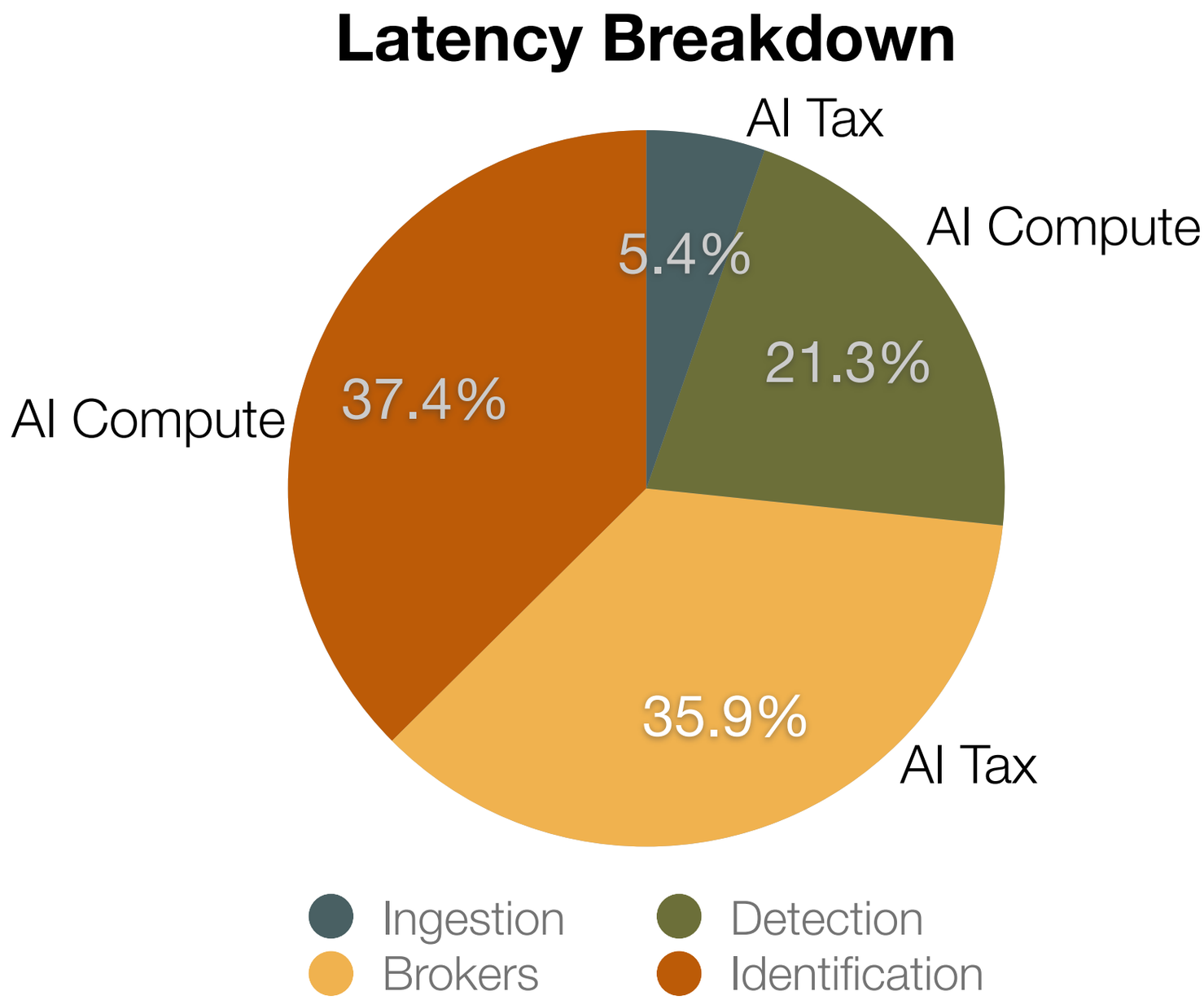
Face Detection Latency



Face Detection Latency



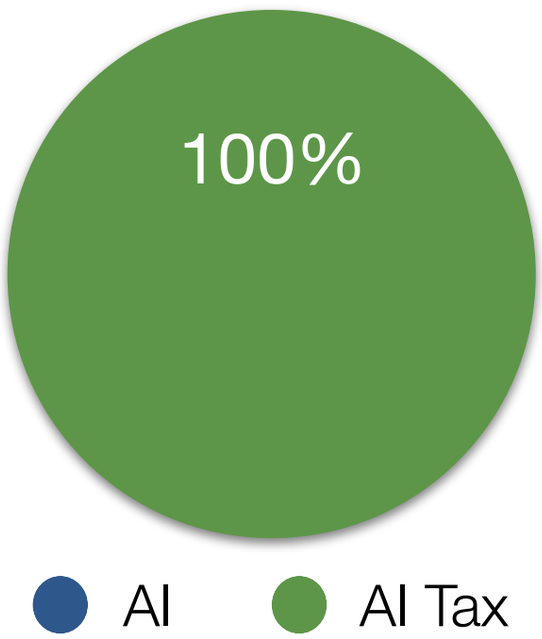
Face Detection Latency



Process Breakdowns

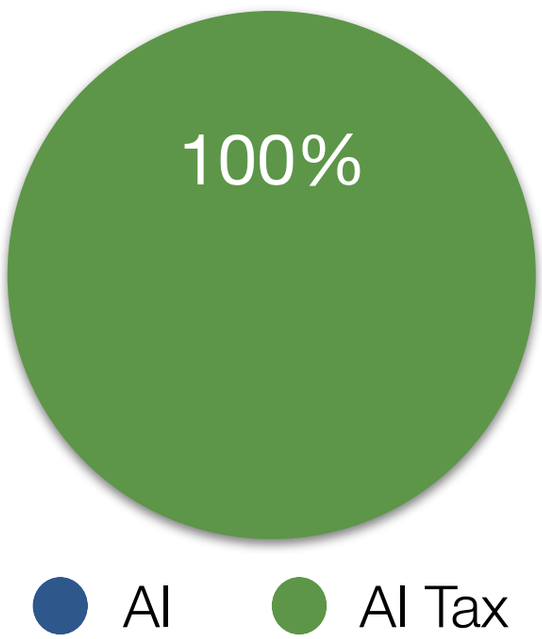
Process Breakdowns

Ingestion

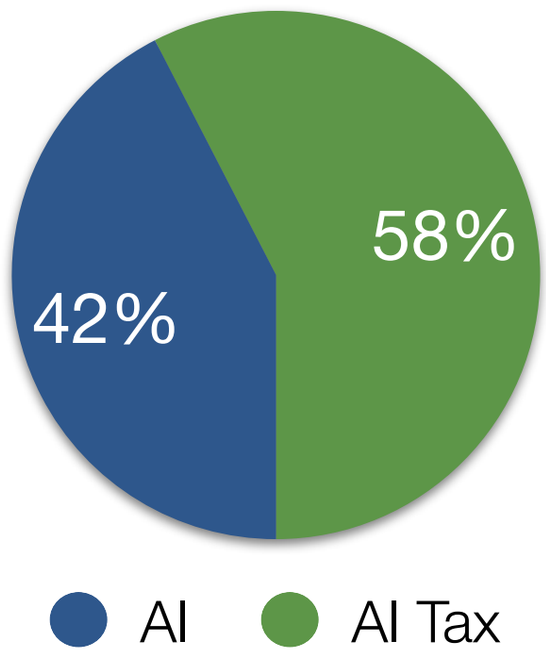


Process Breakdowns

Ingestion

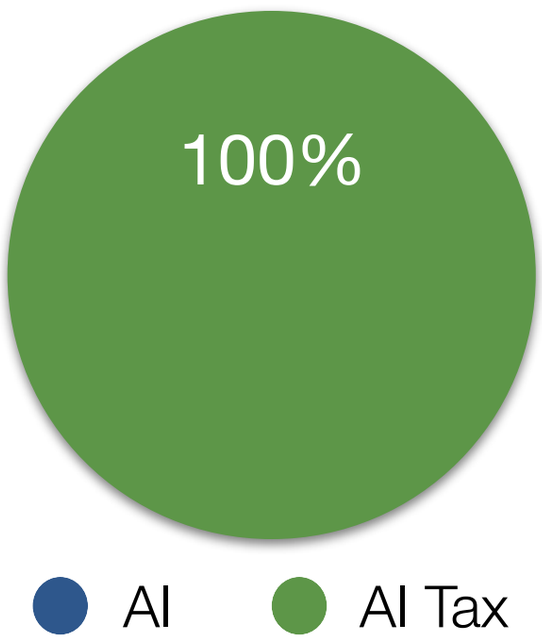


Face Detection

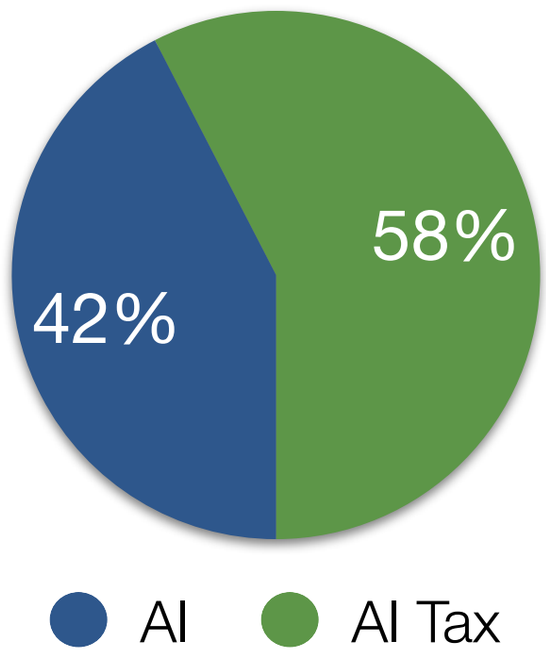


Process Breakdowns

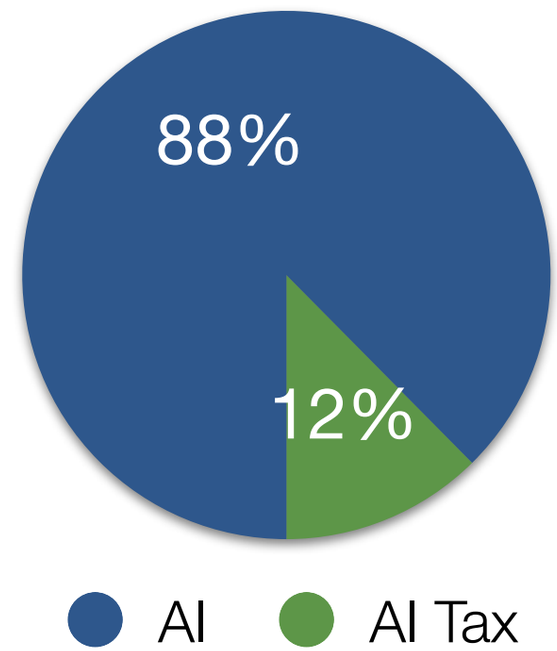
Ingestion



Face Detection



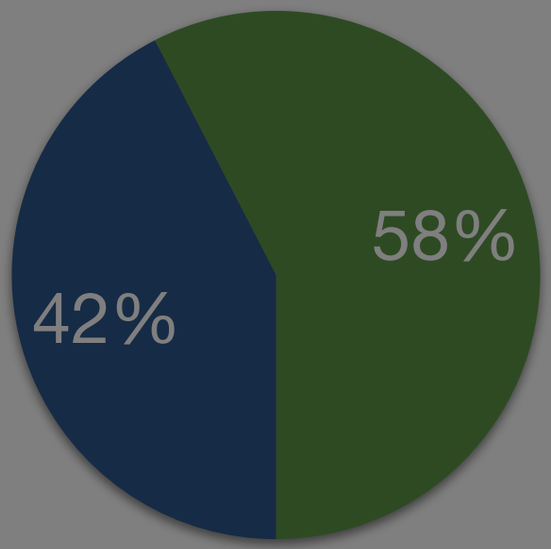
Identification



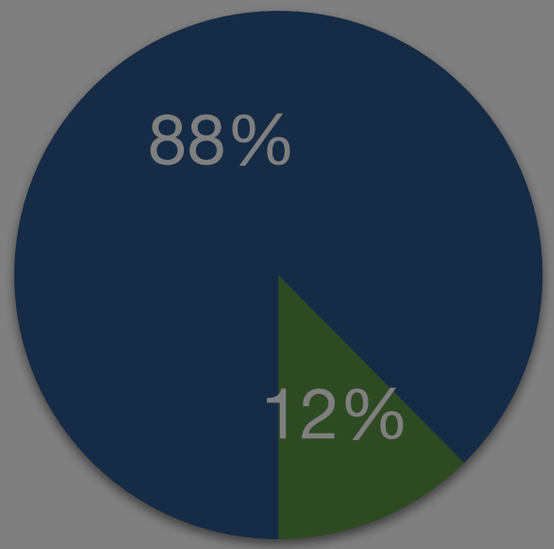
Process Breakdowns

Ingestion

Face Detection

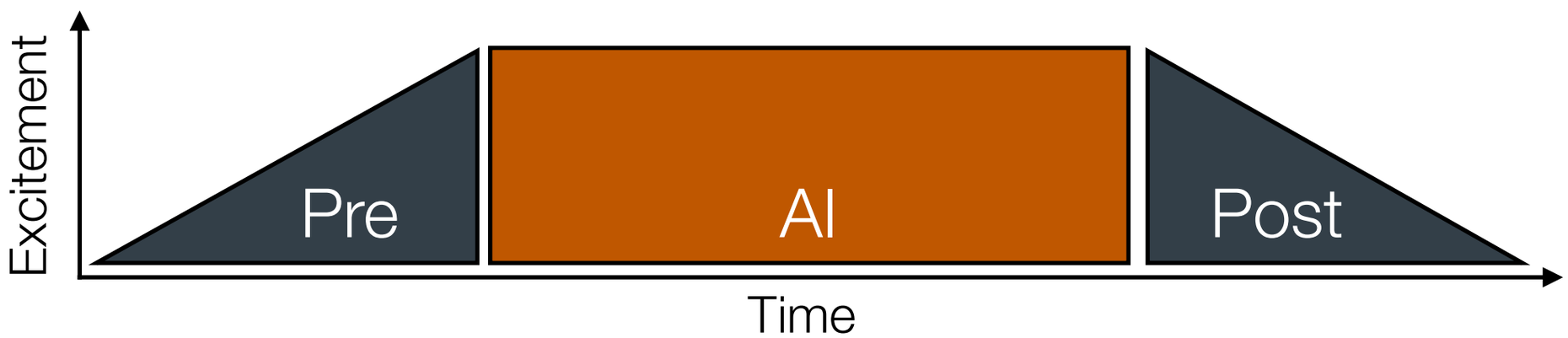


Identification

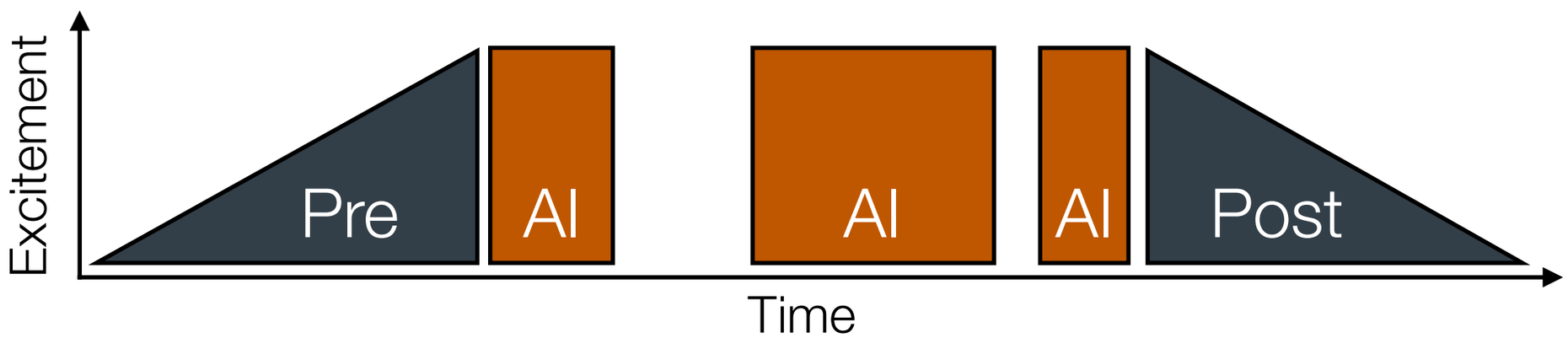


Pre- and post-processing are heavily utilized within stages.

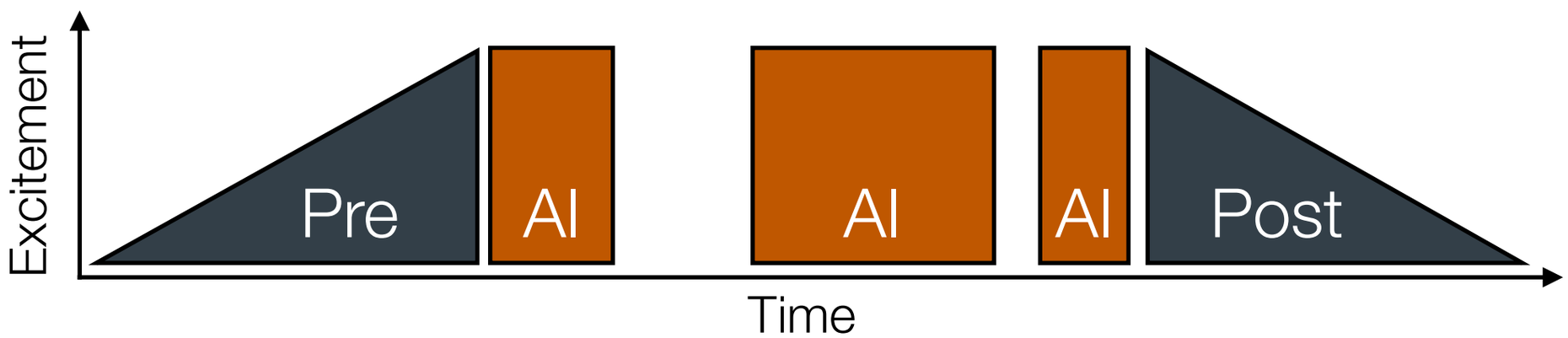
AI Tax



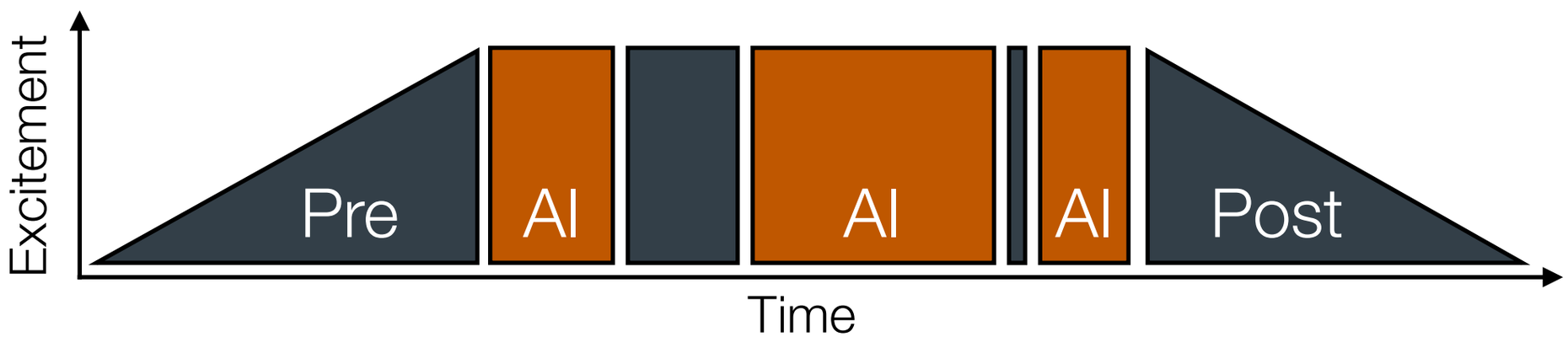
AI Tax



AI Tax



AI Tax

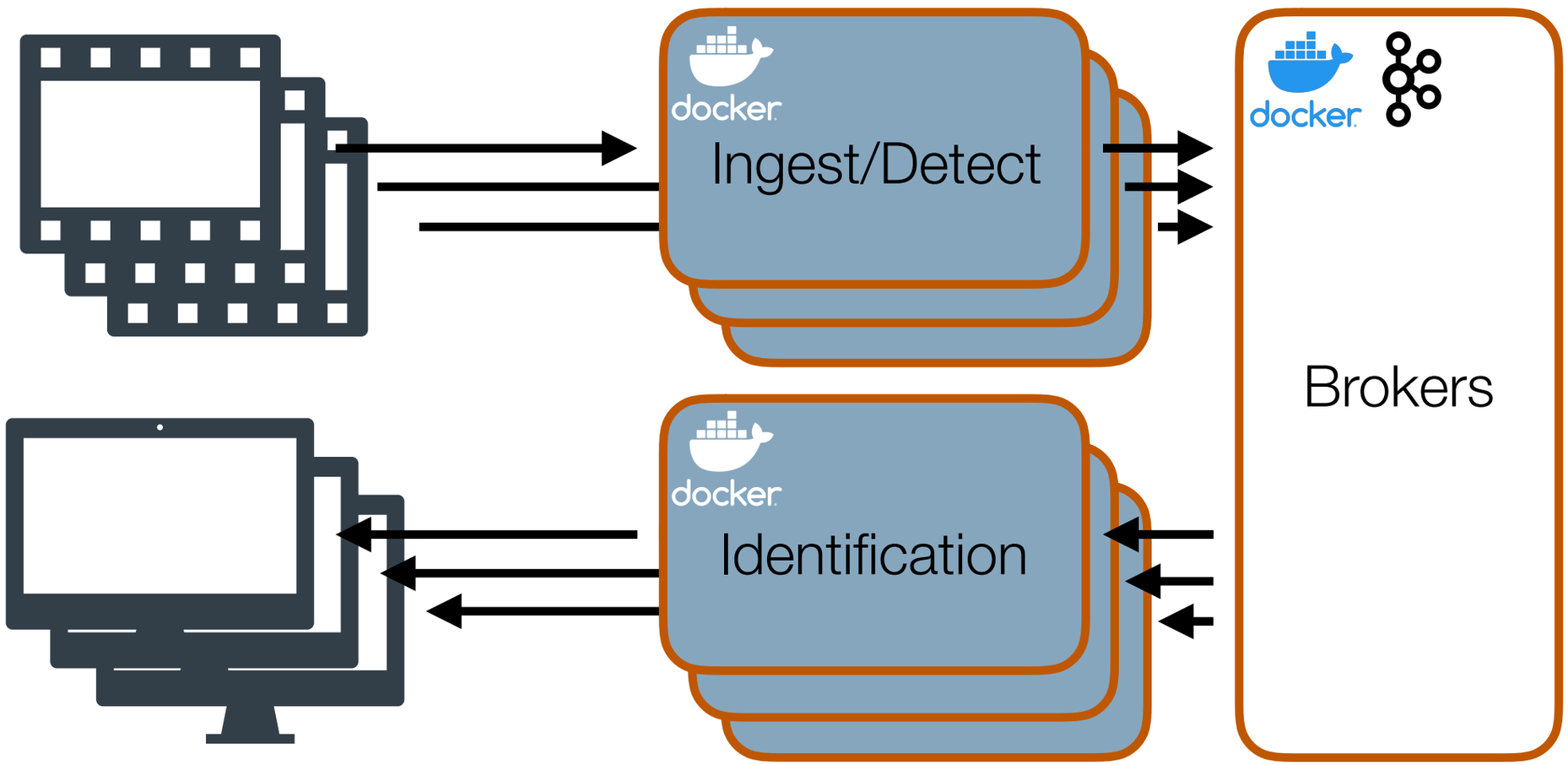


1. AI Tax
    a. Definition
    b. Video Analytics
    c. Analysis
2. AI Acceleration
    a. Emulation Technique
    b. Results
    c. What's Breaking?
3. Optimization
    a. Fixing the Bottleneck
    b. Edge Data Centers
    c. Two Designs
4. Conclusion

AI Acceleration

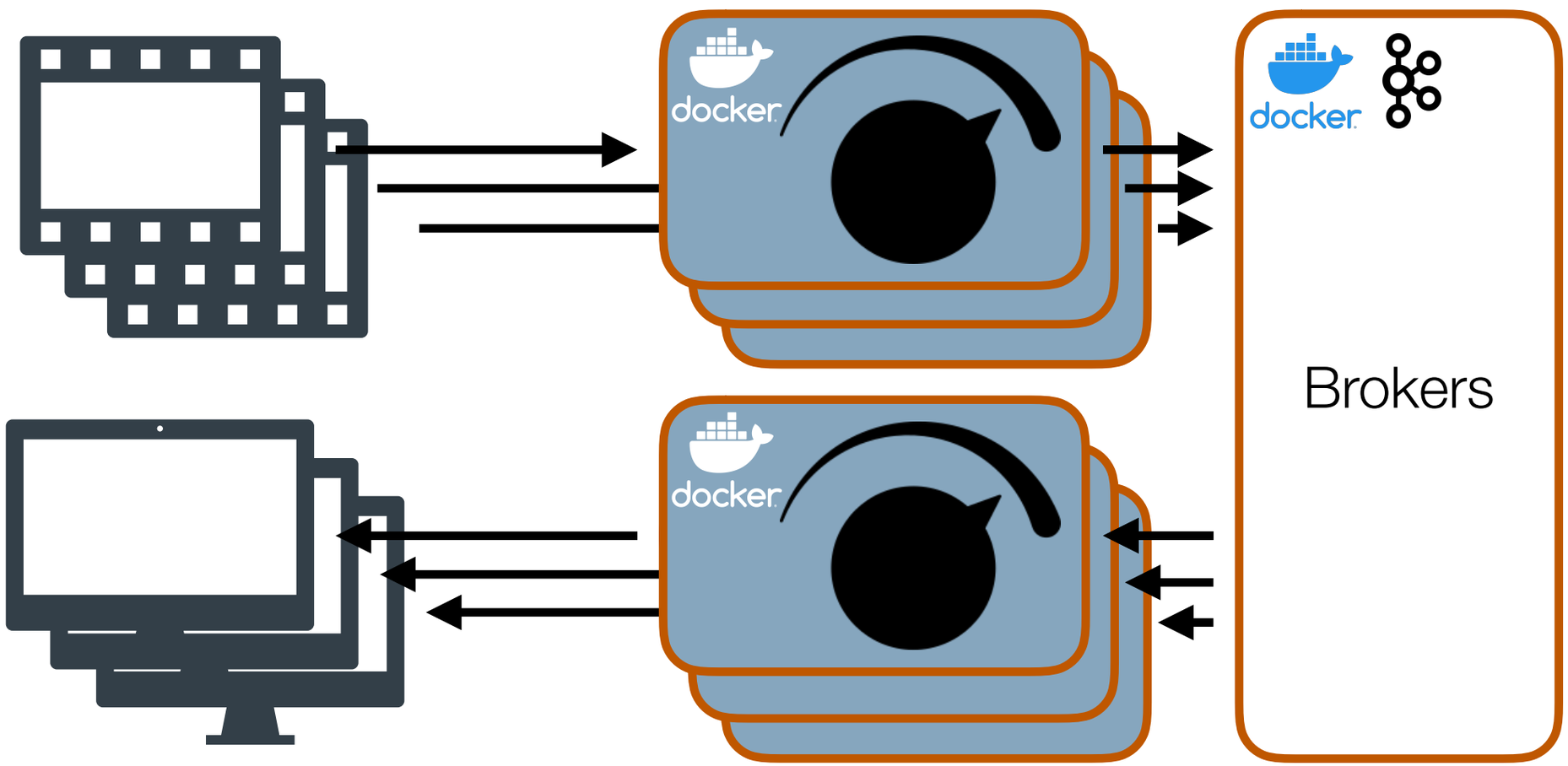
AI Acceleration - Emulation Technique

Acceleration Emulation



Acceleration Emulation

Dial an Accelerator Speed



Acceleration Emulation

```
while True:
    frame = queue.get()
    start = time.time()
    faces = detect_faces(frame)
    end = time.time()
    producer.send(faces)
    size = sys.getsizeof(faces)
    log = {
        'start': start,
        'end': end,
        'size': size
    }
    logger.info(log)
```

Acceleration Emulation

```
while True:
    frame = queue.get()
    start = time.time()
    faces = detect_faces(frame)
    end = time.time()
    producer.send(faces)
    size = sys.getsizeof(faces)
    log = {
        'start': start,
        'end': end,
        'size': size
    }
    logger.info(log)
```

Acceleration Emulation

```
while True:
    frame = queue.get()
    start = time.time()
    time.sleep(avg_time)
    end = time.time()
    producer.send(faces)
    size = sys.getsizeof(faces)
    log = {
        'start': start,
        'end': end,
        'size': size
    }
    logger.info(log)
```

Acceleration Emulation

```
while True:
    frame = queue.get()
    start = time.time()
    time.sleep(avg_time)
    end = time.time()
    producer.send(faces)
    size = sys.getsizeof(faces)
    log = {
        'start': start,
        'end': end,
        'size': size
    }
    logger.info(log)
```

Acceleration Emulation

```
while True:
    frame = queue.get()
    start = time.time()
    time.sleep(avg_time)
    end = time.time()
    producer.send(os.urandom(avg_size))
    size = avg_size
    log = {
        'start': start,
        'end': end,
        'size': size
    }
    logger.info(log)
```

Acceleration Emulation

```
while True:
    frame = queue.get()
    start = time.time()
    time.sleep(avg_time)
    end = time.time()
    producer.send(os.urandom(avg_size))
    size = avg_size
    log = {
        'start': start,
        'end': end,
        'size': size
    }
    logger.info(log)
```

Acceleration Emulation

```
while True:
    frame = queue.get()
    start = time.time()
    time.sleep(avg_time)
    end = time.time()
    producer.send(os.urandom(avg_size))
    size = avg_size
    log = {
        'start': start,
        'end': end,
        'size': size
    }
    logger.info(log)
```

Acceleration Emulation

```
while True:
    frame = queue.get()
    start = time.time()
    time.sleep(avg_time)
    end = time.time()
    producer.send(os.urandom(avg_size))
    size = avg_size
    log = {
        'start': start,
        'end': end,
        'size': size
    }
    logger.info(log)
```

Acceleration Emulation

```
while True:
    frame = queue.get()
    start = time.time()
    time.sleep(avg_time/speedup)
    end = time.time()
    producer.send(os.urandom(avg_size))
    size = avg_size
    log = {
        'start': start,
        'end': end,
        'size': size
    }
    logger.info(log)
```

Acceleration Emulation

```
while True:
    frame = queue.get()
    start = time.time()
    time.sleep(avg_time/speedup)
    end = time.time()
    producer.send(os.urandom(avg_size))
    size = avg_size
    log = {
        'start': start,
        'end': end,
        'size': size
    }
    logger.info(log)
```

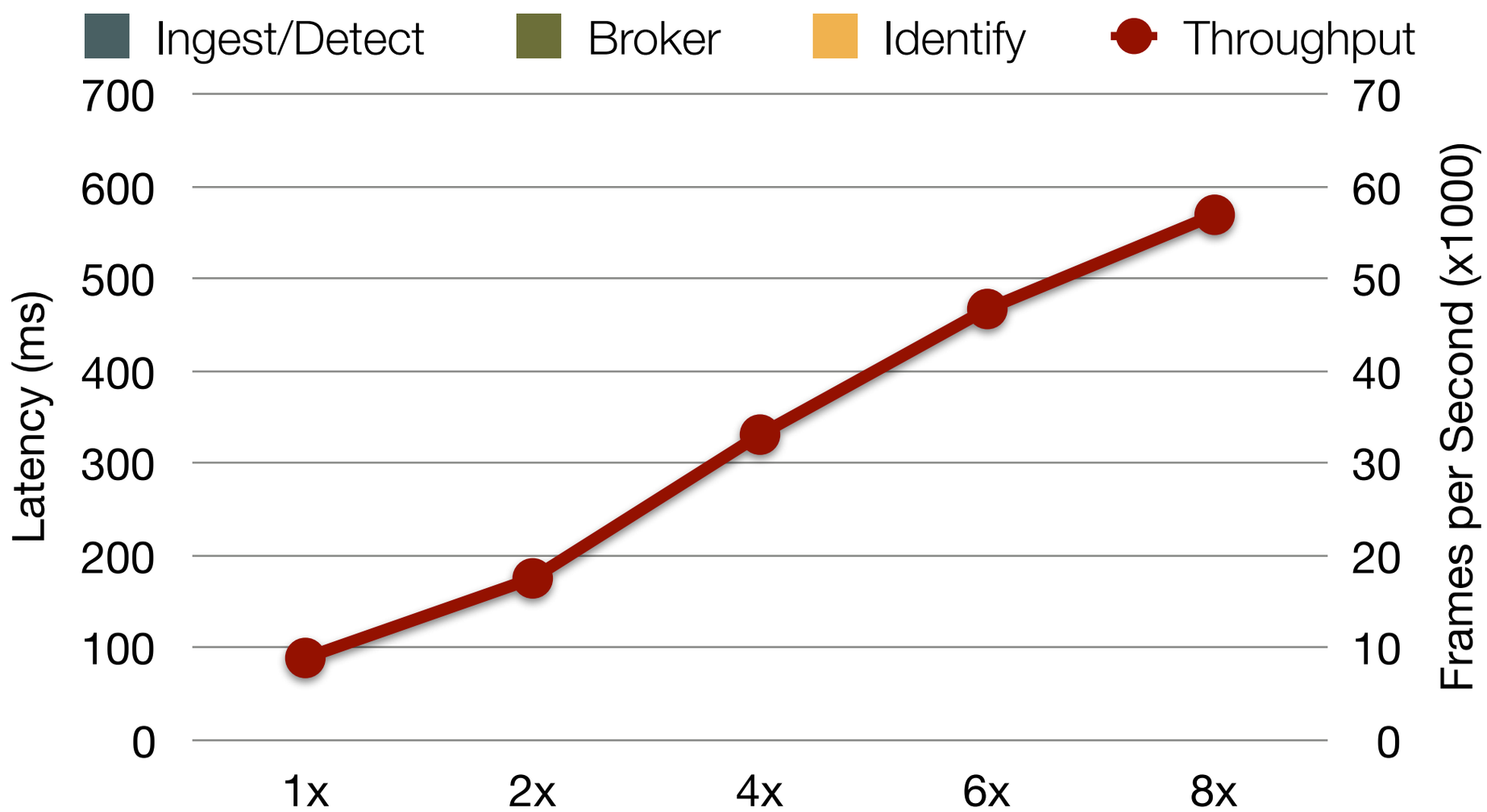
With faster processing, we feed frames into the system faster to maximize throughput

AI Acceleration - Results

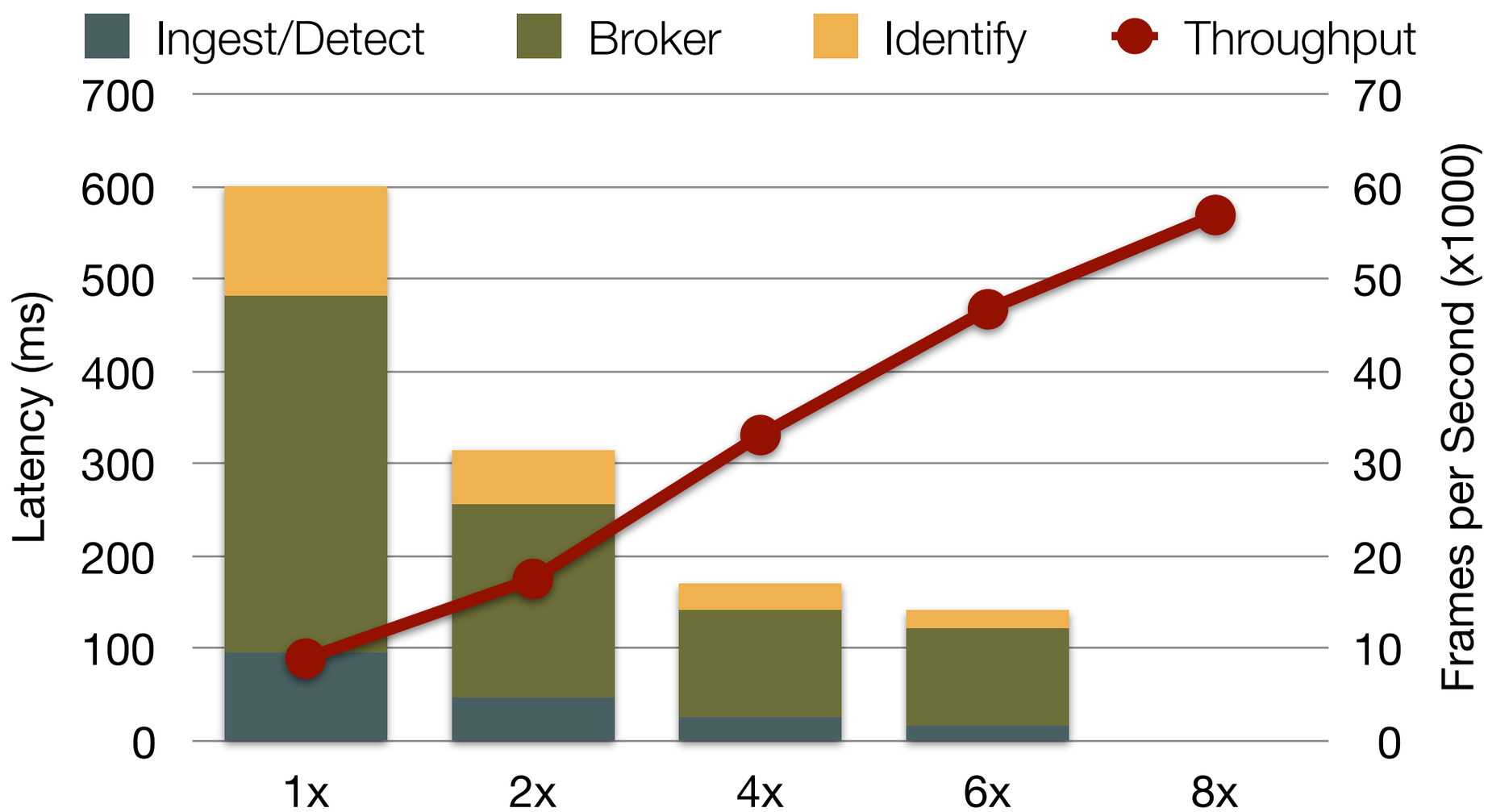
Accelerated AI:
Reduced Latency and Increased Throughput

Accelerated AI:
Reduced Latency and Increased Throughput

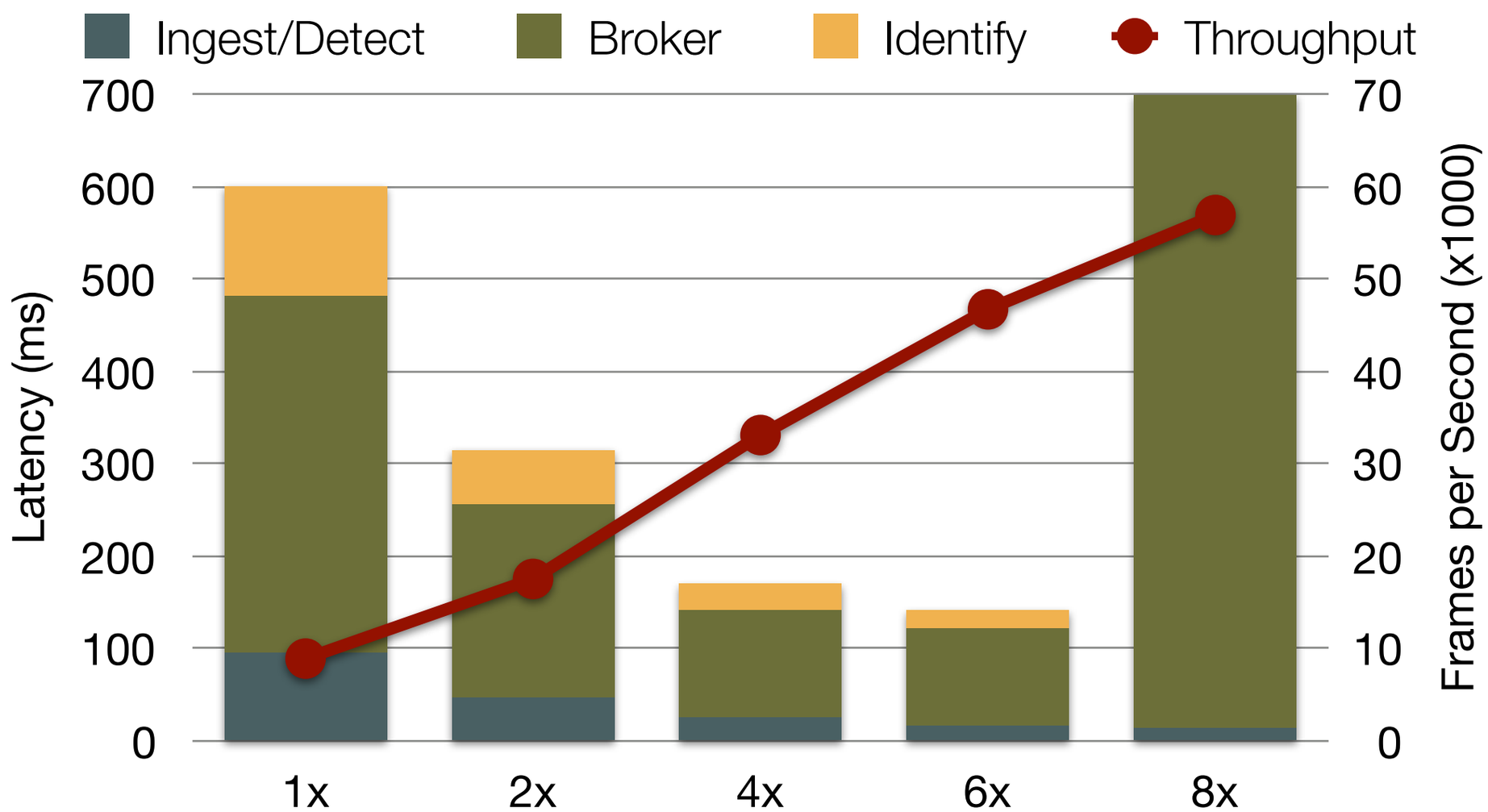
Accelerated AI:
Reduced Latency and Increased Throughput



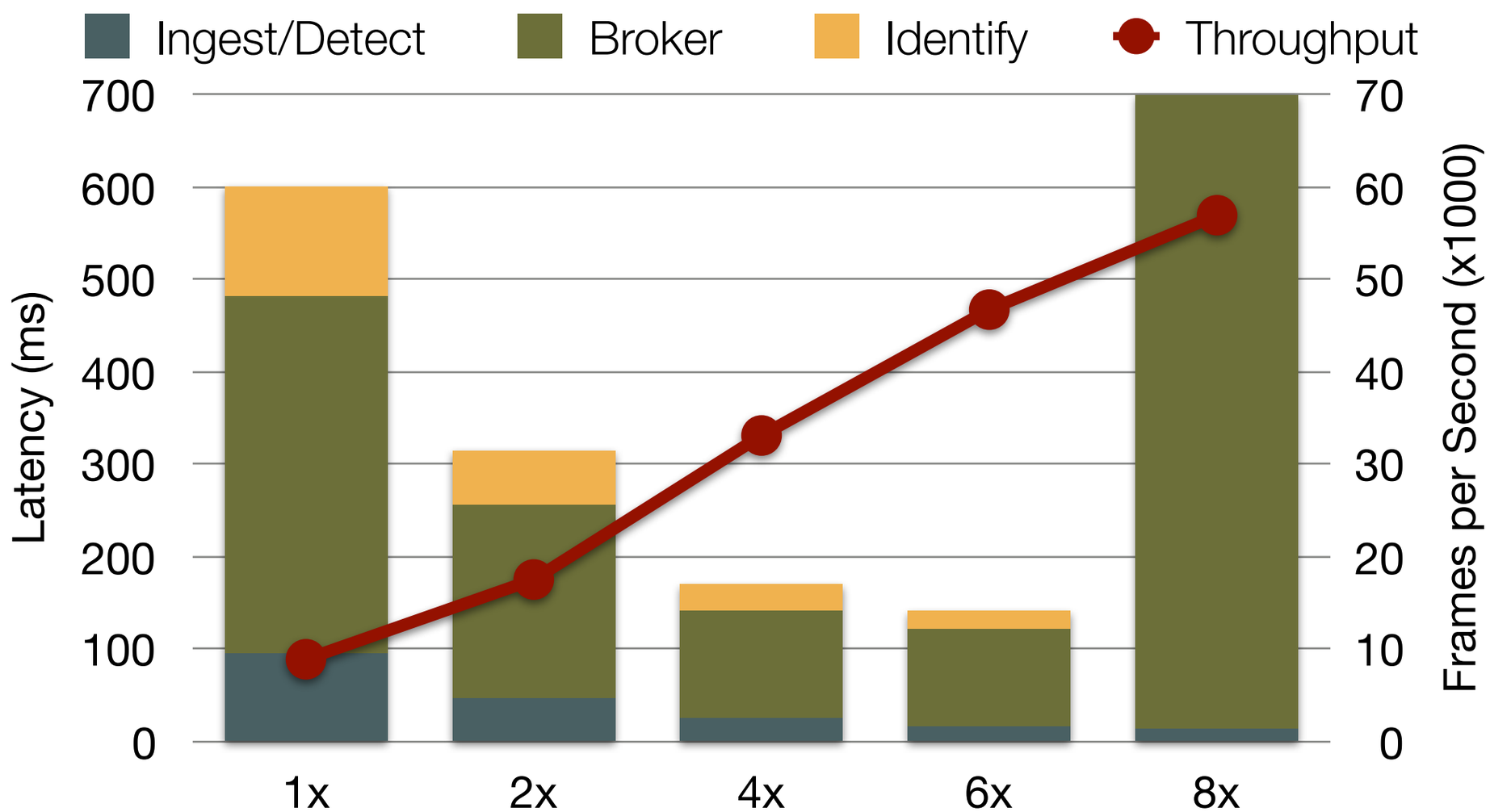
Accelerated AI:
Reduced Latency and Increased Throughput



Accelerated AI:
Reduced Latency and Increased Throughput



Accelerated AI:
Reduced Latency and Increased Throughput



At 8x speedup, the average latency goes to infinity.

The longer the experiment runs, the greater the latency.

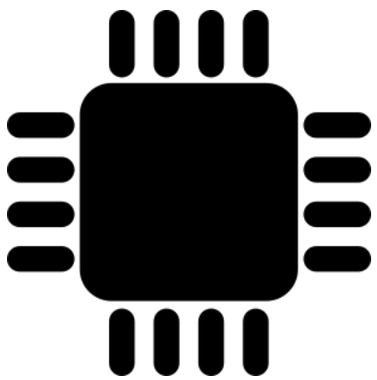
1. AI Tax
    a. Definition
    b. Video Analytics
    c. Analysis
2. AI Acceleration
    a. Emulation Technique
    b. Results
    c. What's Breaking?
3. Optimization
    a. Fixing the Bottleneck
    b. Edge Data Centers
    c. Two Designs
4. Conclusion

AI Acceleration - What's Breaking?

Three Big Systems

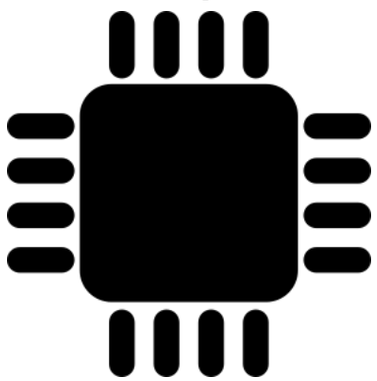
Three Big Systems

Compute

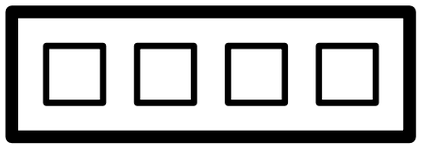


Three Big Systems

Compute

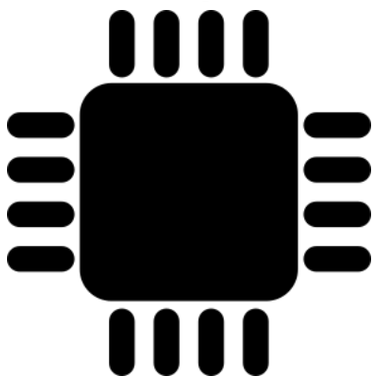


Network

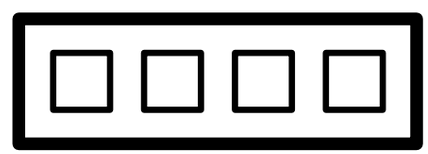


Three Big Systems

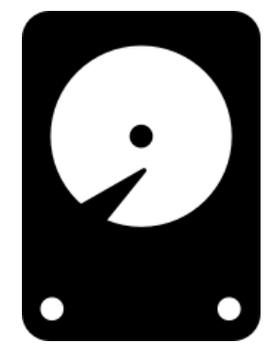
Compute



Network



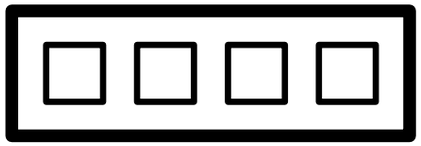
Storage



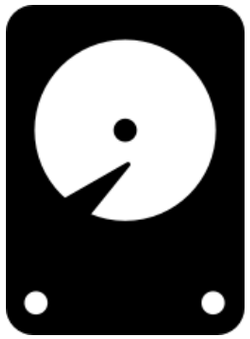
Three Big Systems

Compute

Network



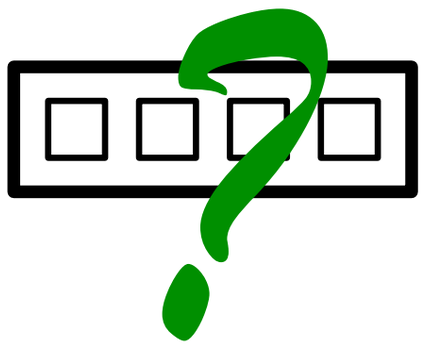
Storage



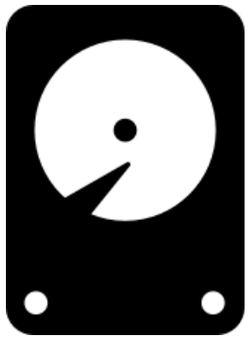
Three Big Systems

Compute

Network



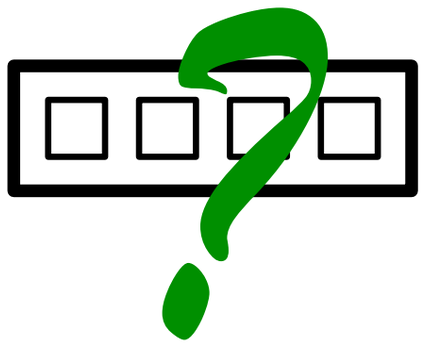
Storage



Three Big Systems

Compute

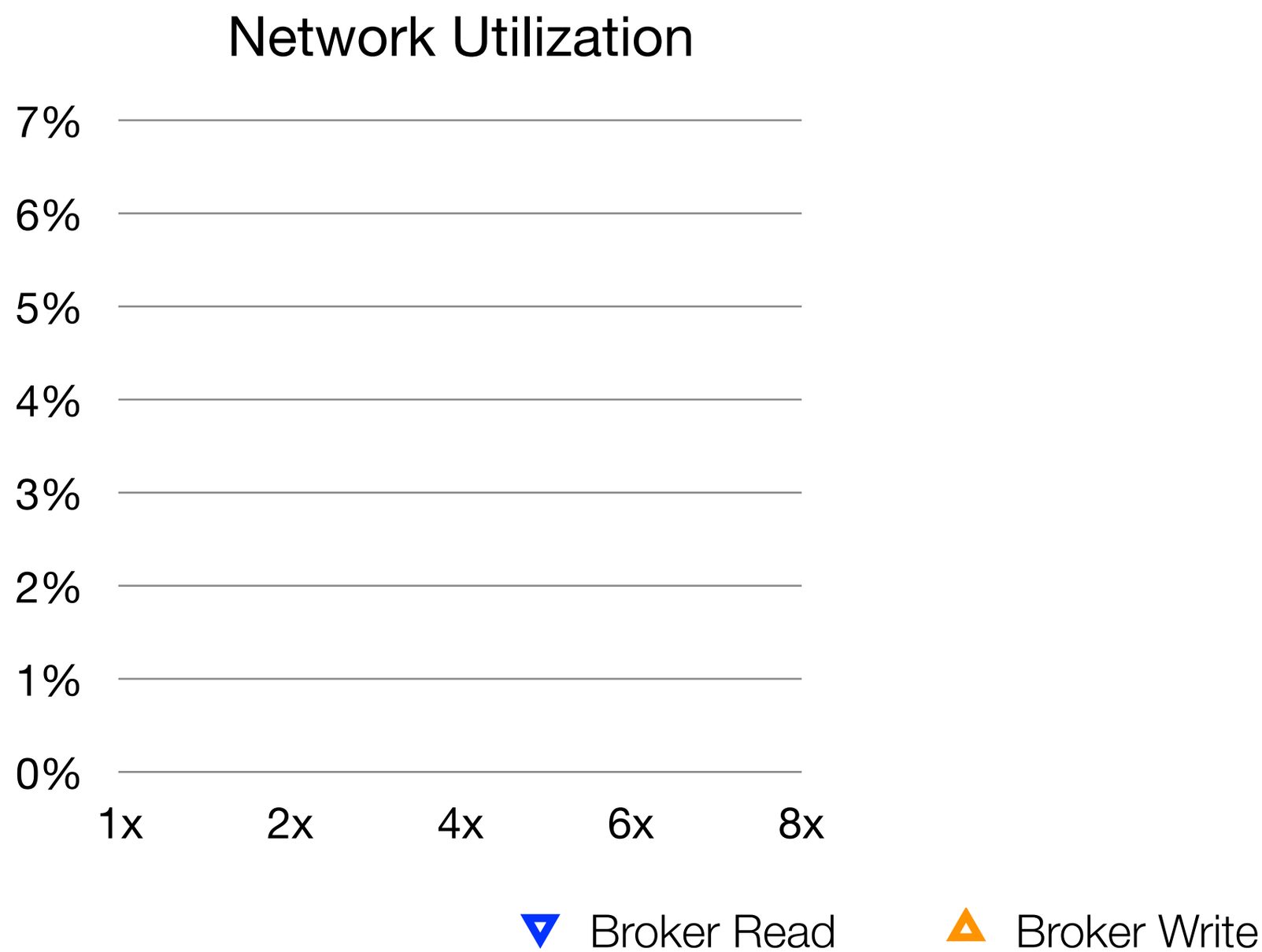
Network



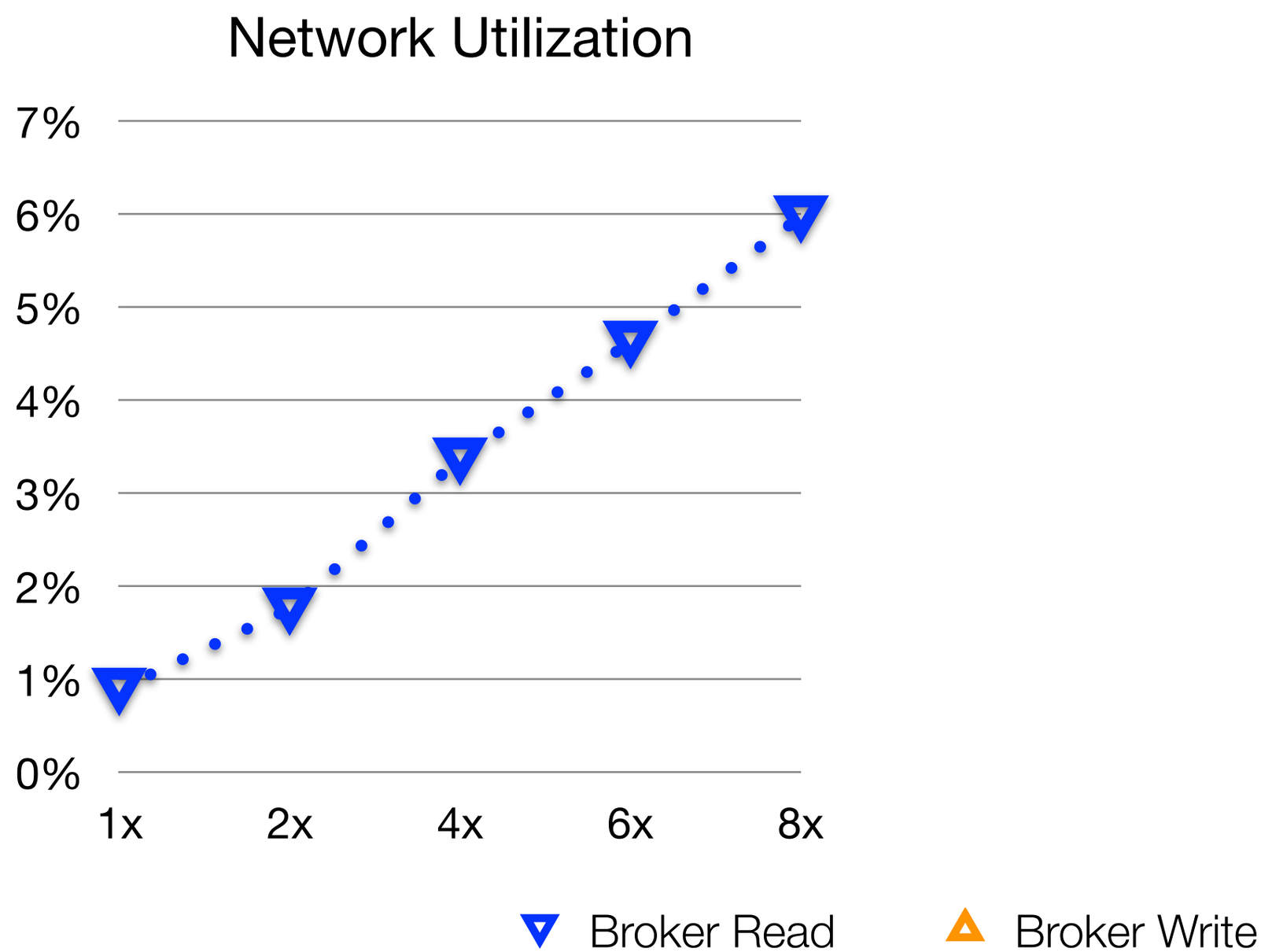
Storage

Explaining the Bottleneck

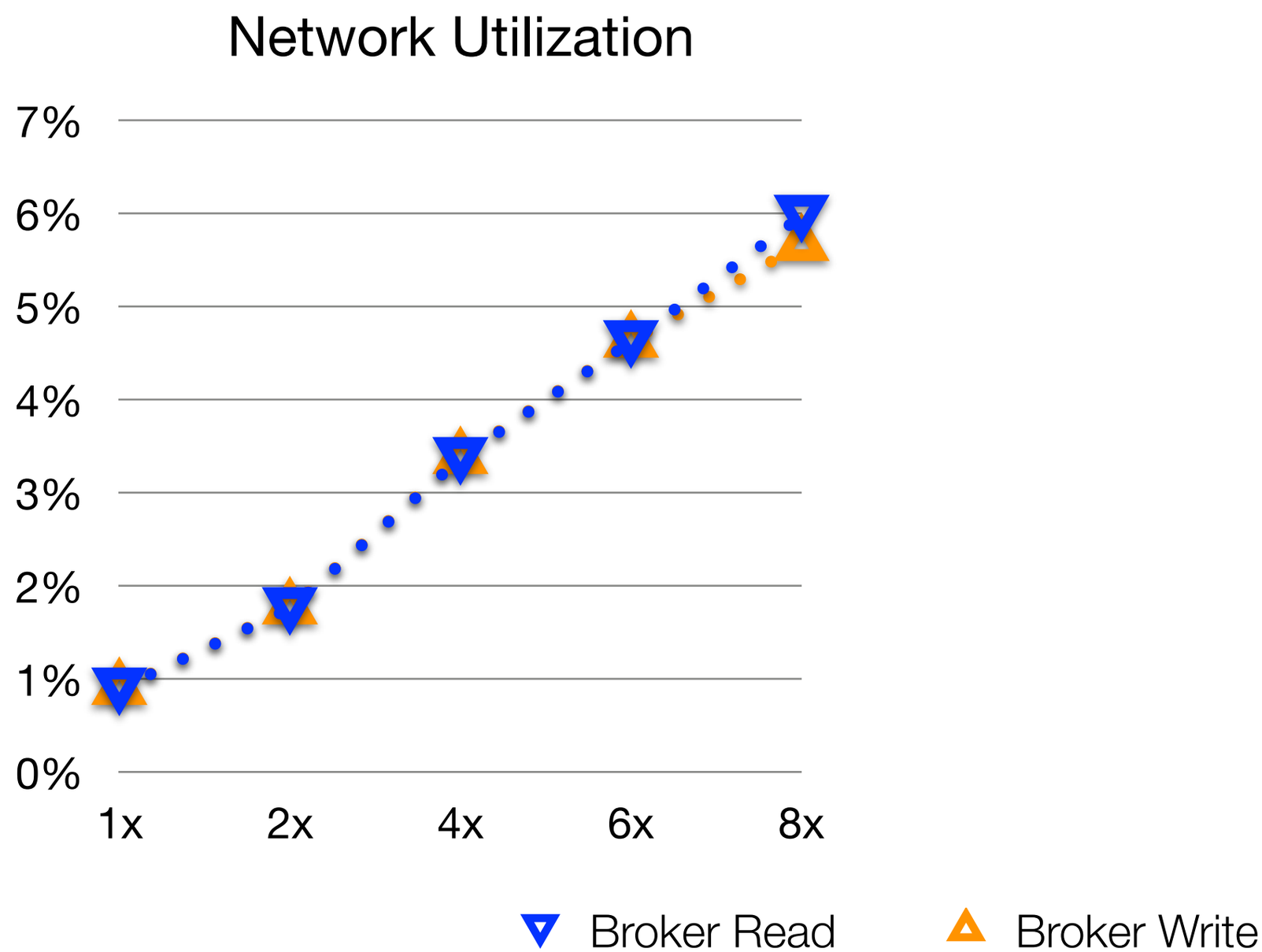
Explaining the Bottleneck



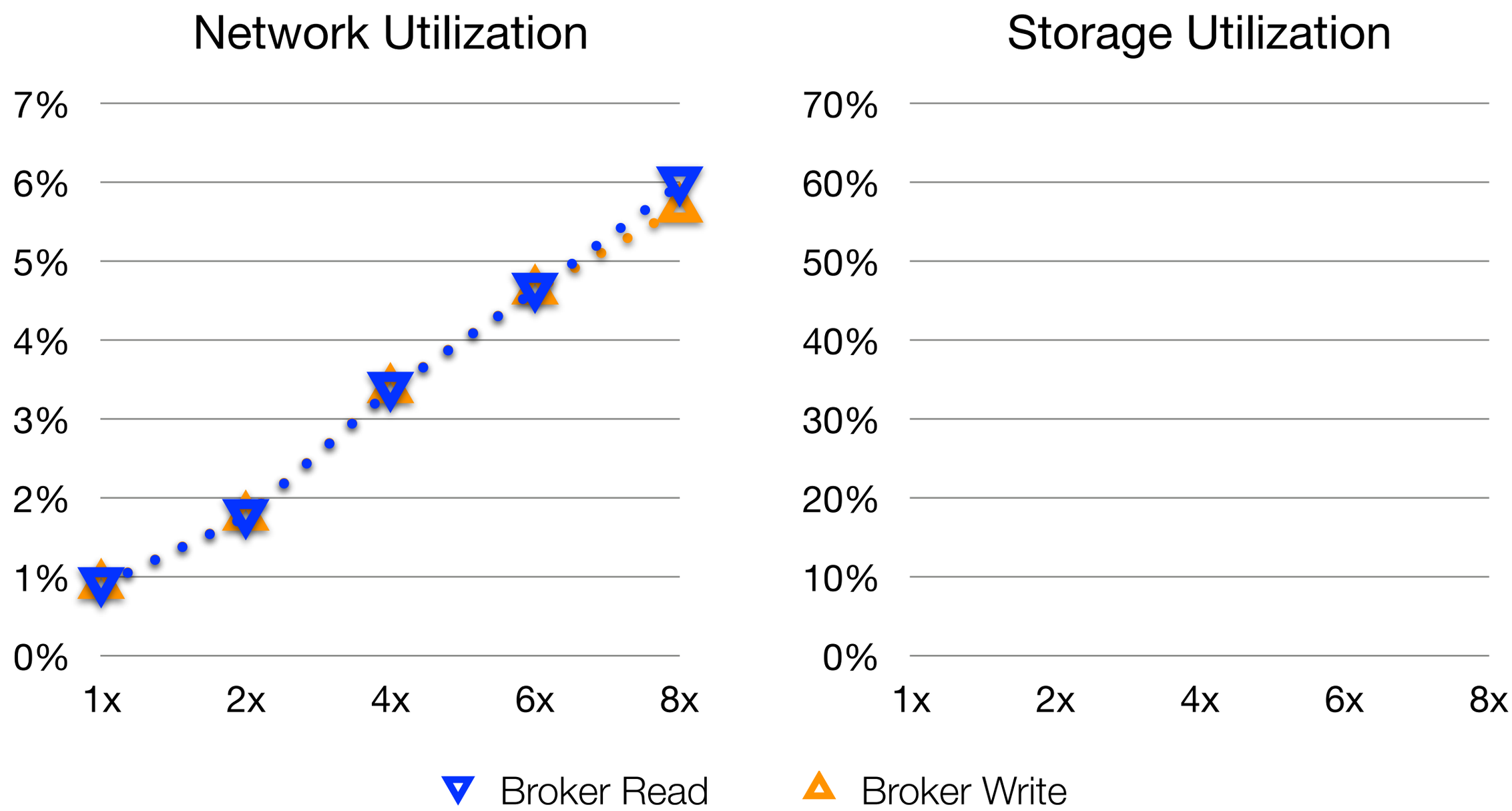
Explaining the Bottleneck



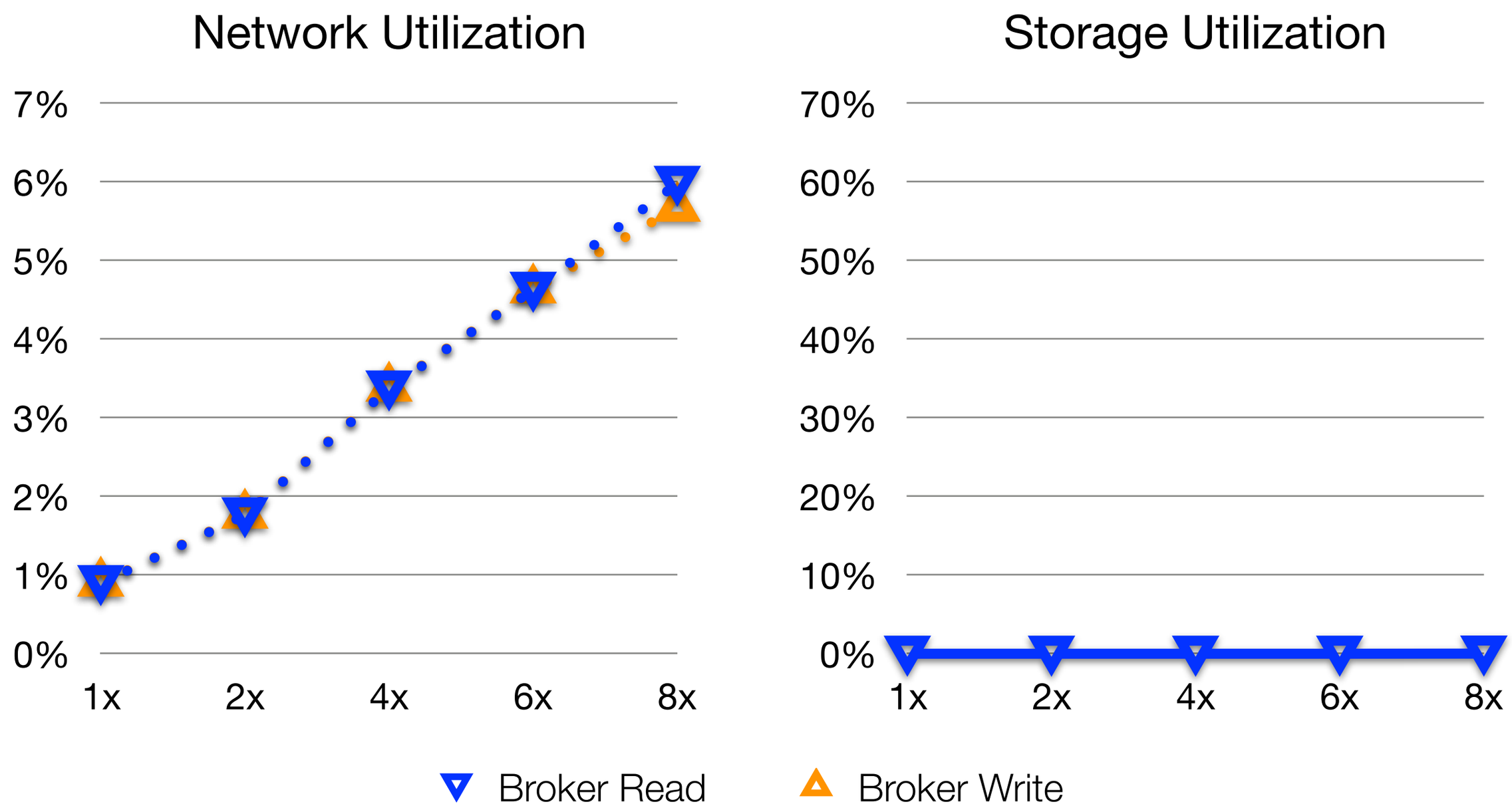
Explaining the Bottleneck



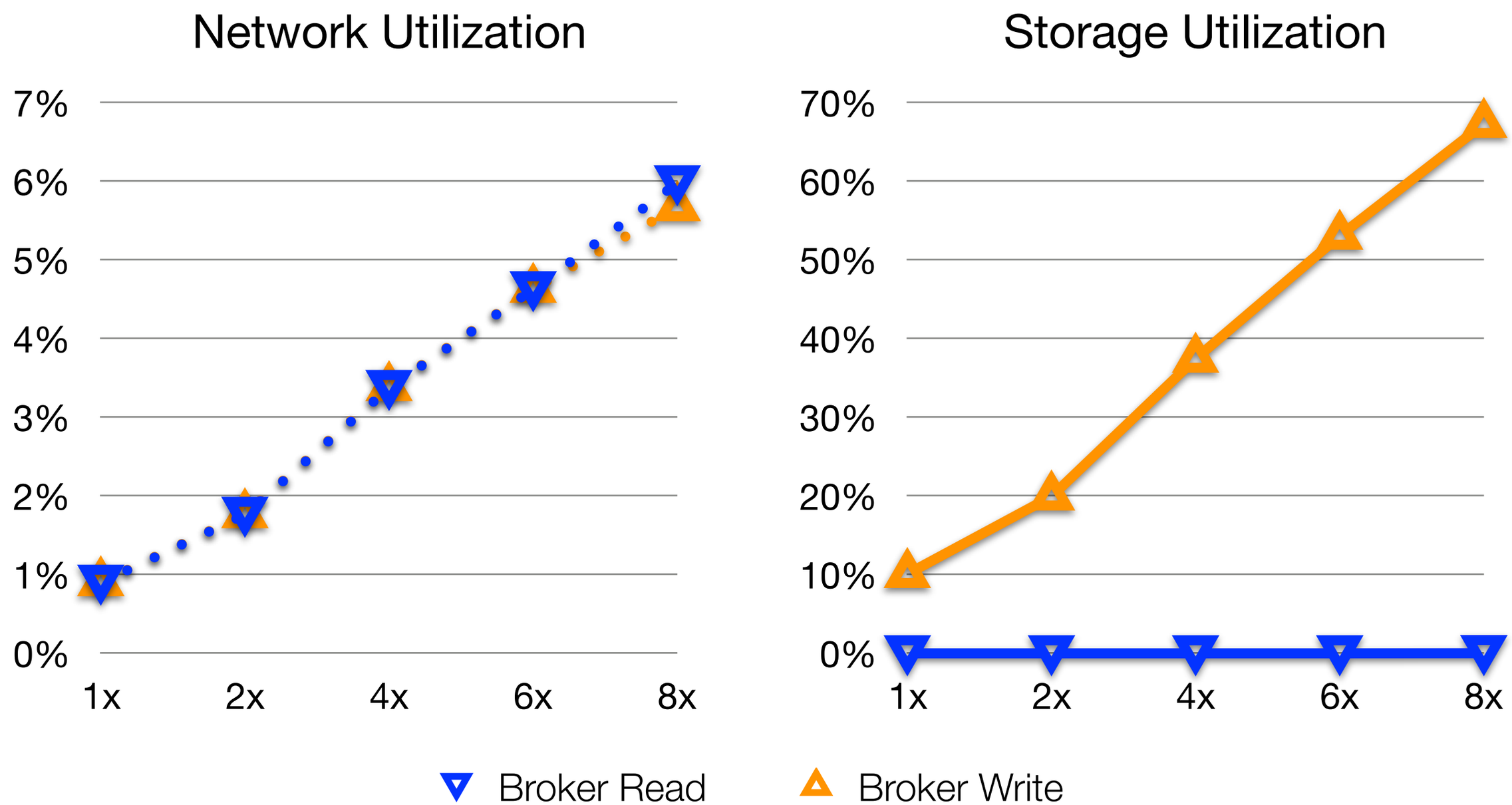
Explaining the Bottleneck



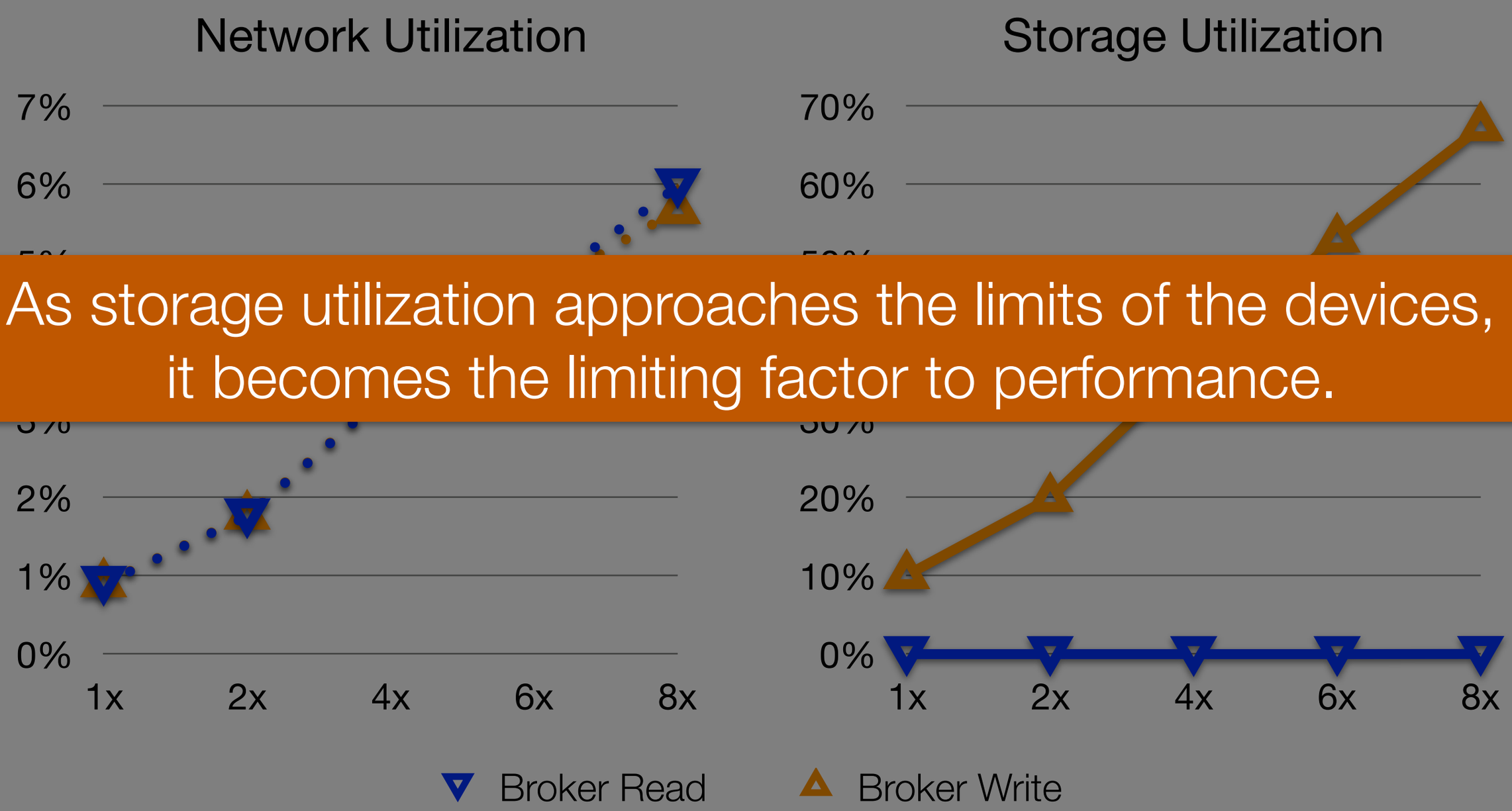
Explaining the Bottleneck



Explaining the Bottleneck



Explaining the Bottleneck



As storage utilization approaches the limits of the devices, it becomes the limiting factor to performance.

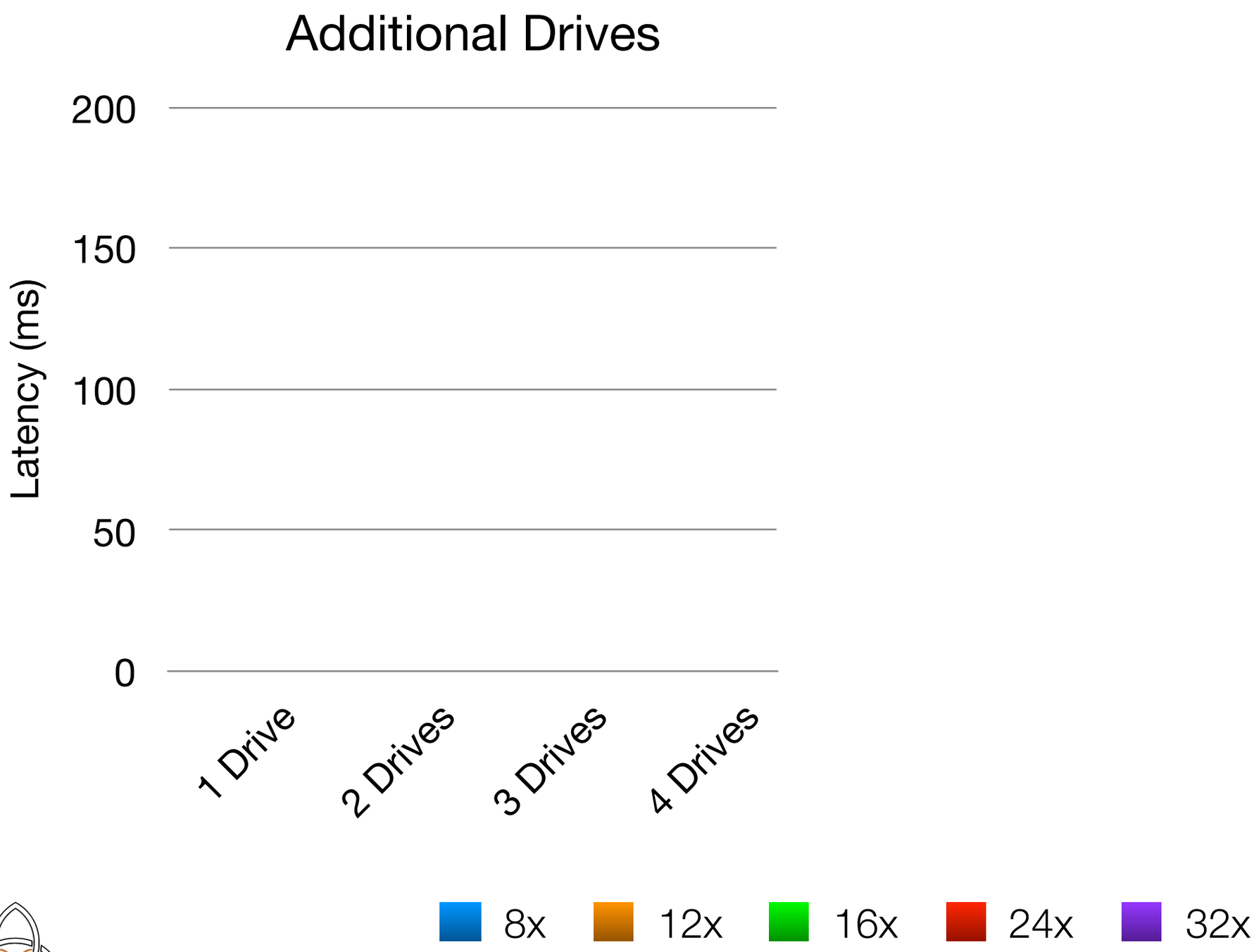
1. AI Tax
    a. Definition
    b. Video Analytics
    c. Analysis
2. AI Acceleration
    a. Emulation Technique
    b. Results
    c. What's Breaking?
3. Optimization
    a. Fixing the Bottleneck
    b. Edge Data Centers
    c. Two Designs
4. Conclusion

Optimization

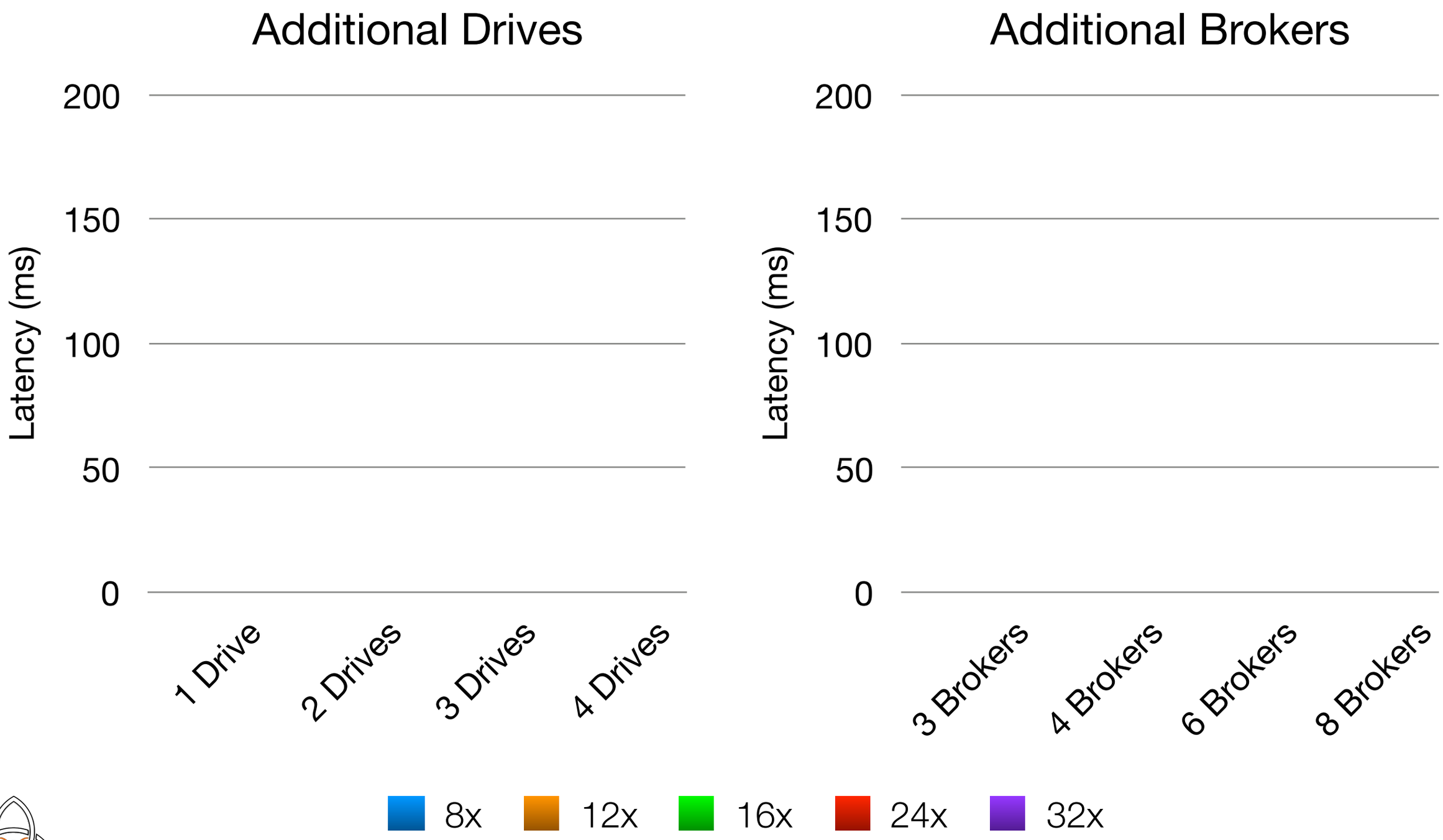
Optimization - Fixing the Bottleneck

Fixing the Bottleneck

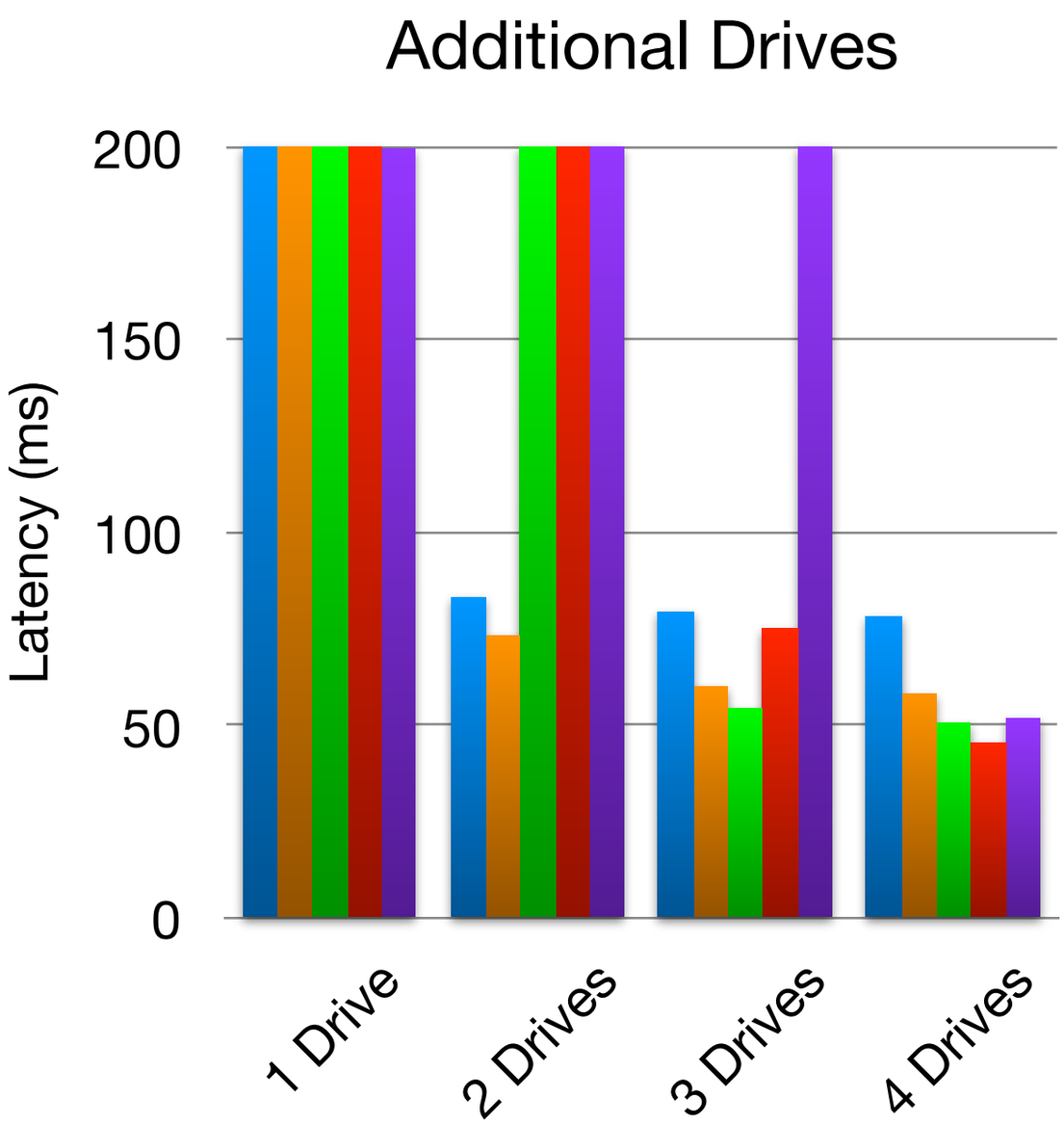
Fixing the Bottleneck

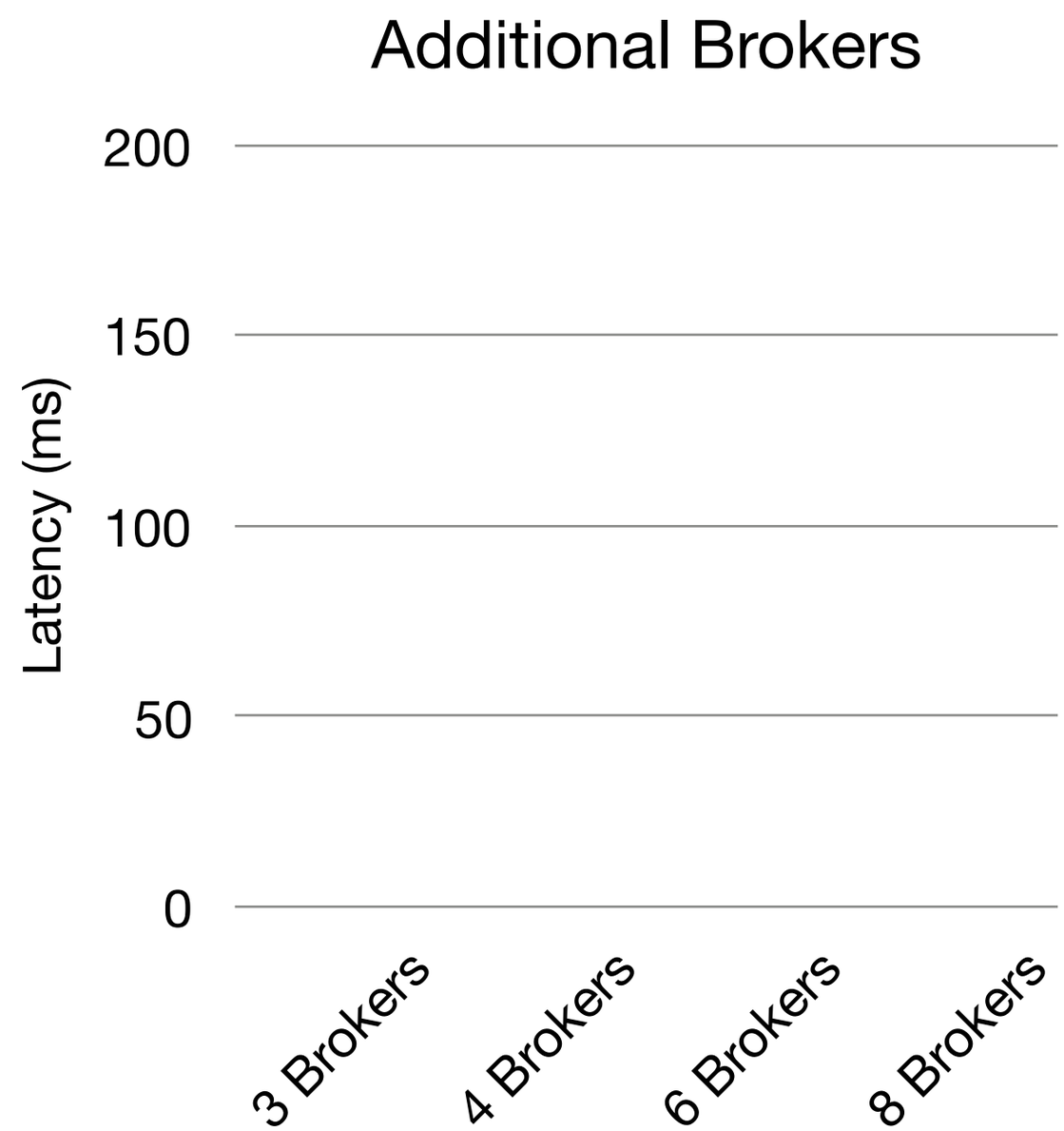


Fixing the Bottleneck



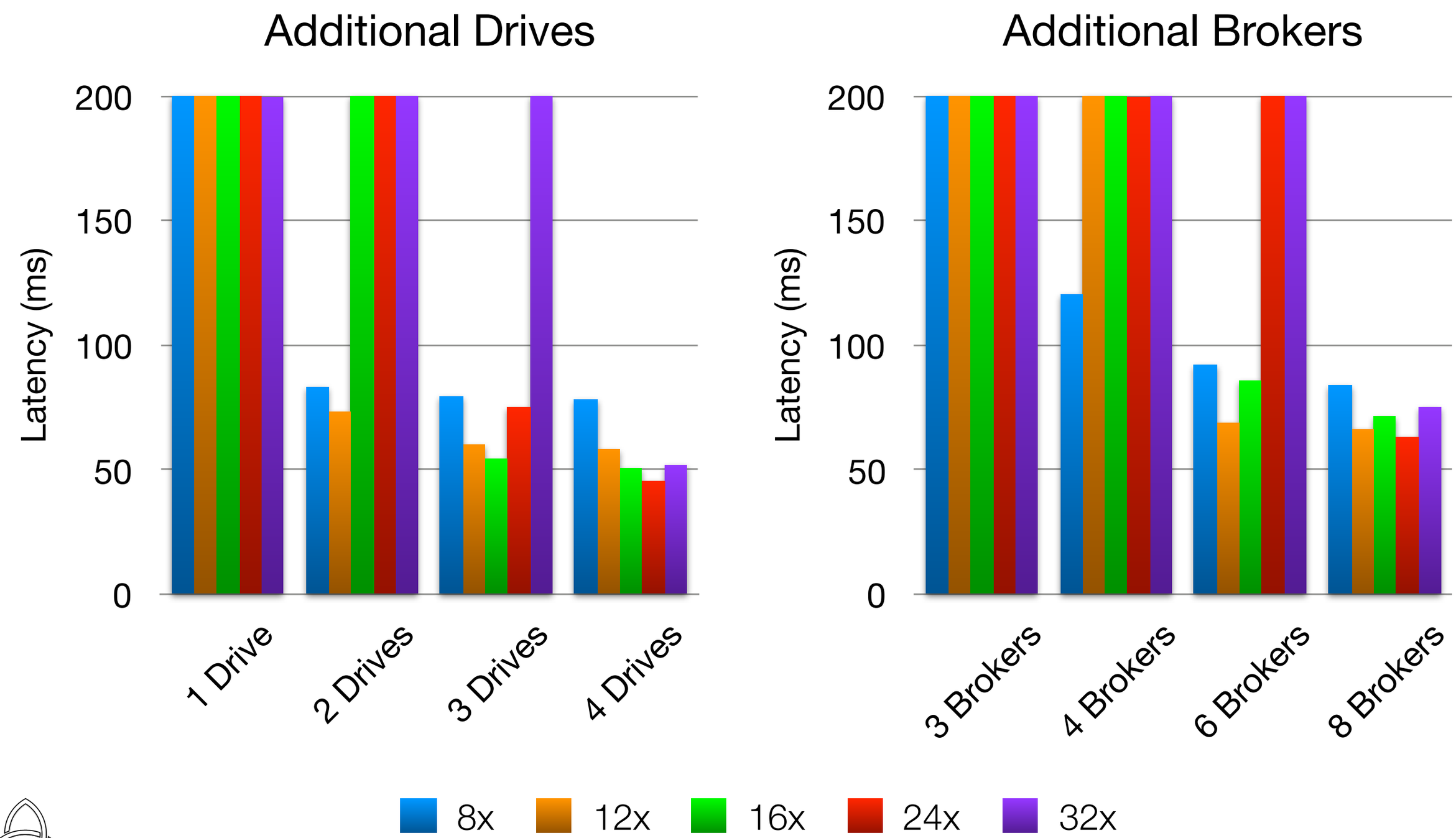
Fixing the Bottleneck





8x 12x 16x 24x 32x

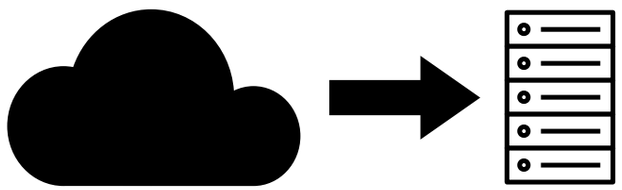
Fixing the Bottleneck



Optimization - Edge Data Centers

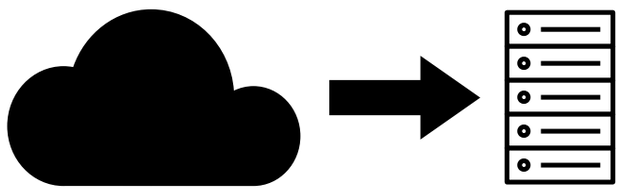
Advantages of an Edge Data Center

Advantages of an Edge Data Center

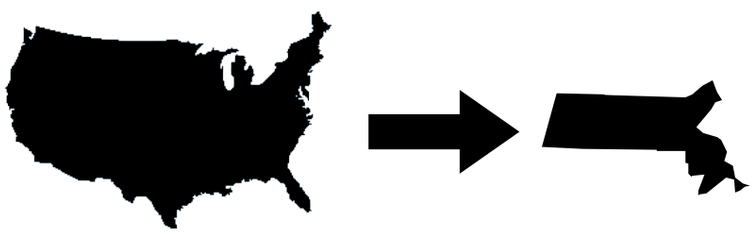


Smaller corporations are finding edge data centers more economical than the cloud.

Advantages of an Edge Data Center

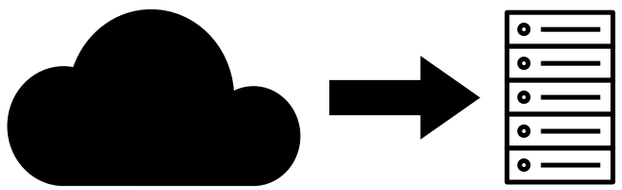


Smaller corporations are finding edge data centers more economical than the cloud.

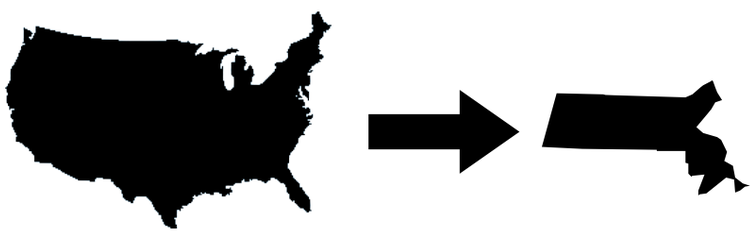


Edge data centers offer lower latency by serving local users.

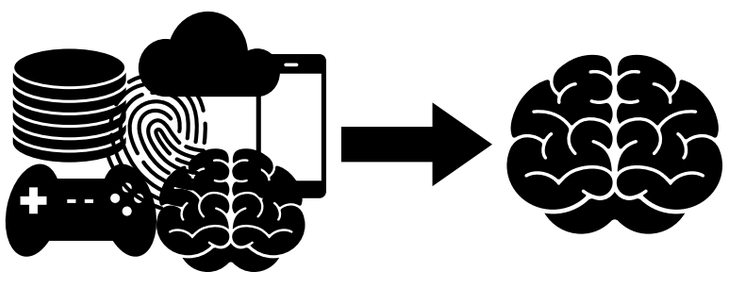
Advantages of an Edge Data Center



Smaller corporations are finding edge data centers more economical than the cloud.



Edge data centers offer lower latency by serving local users.



Edge data centers can be built to target a specific application domain.

Sources https://www.networkworld.com/article/2926448/7-key-criteria-for-defining-edge-data-centers.html
http://blog.cushwake.com/americas/life-on-the-edge-the-new-normal-for-data-centers.html
https://www.vxchnge.com/blog/what-is-an-edge-data-center

1. AI Tax
   a. Definition
   b. Video Analytics
   c. Analysis
2. AI Acceleration
   a. Emulation Technique
   b. Results
   c. What's Breaking?
3. Optimization
   a. Fixing the Bottleneck
   b. Edge Data Centers
   c. Two Designs
4. Conclusion

Optimization - Two Designs

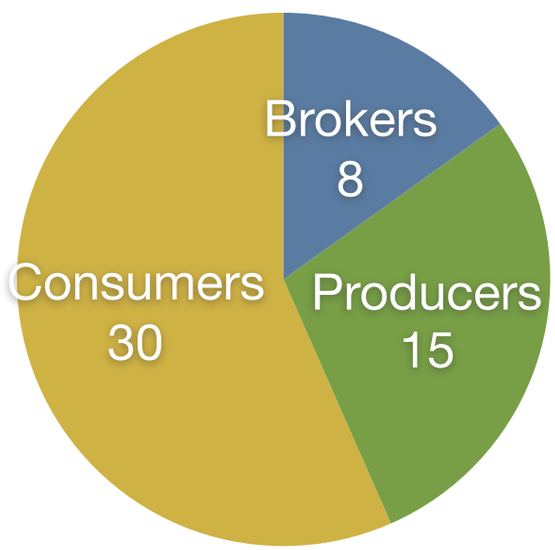
Edge Data Center Node Allocation

Edge Data Center
Node Allocation

We need to allocate enough brokers to handle 32x speedup.

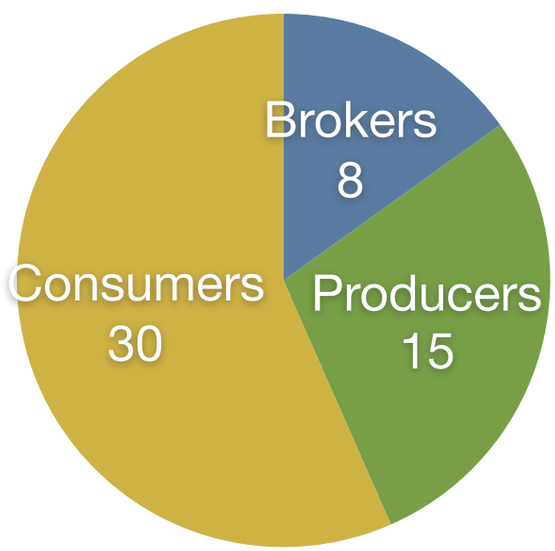
Edge Data Center
Node Allocation

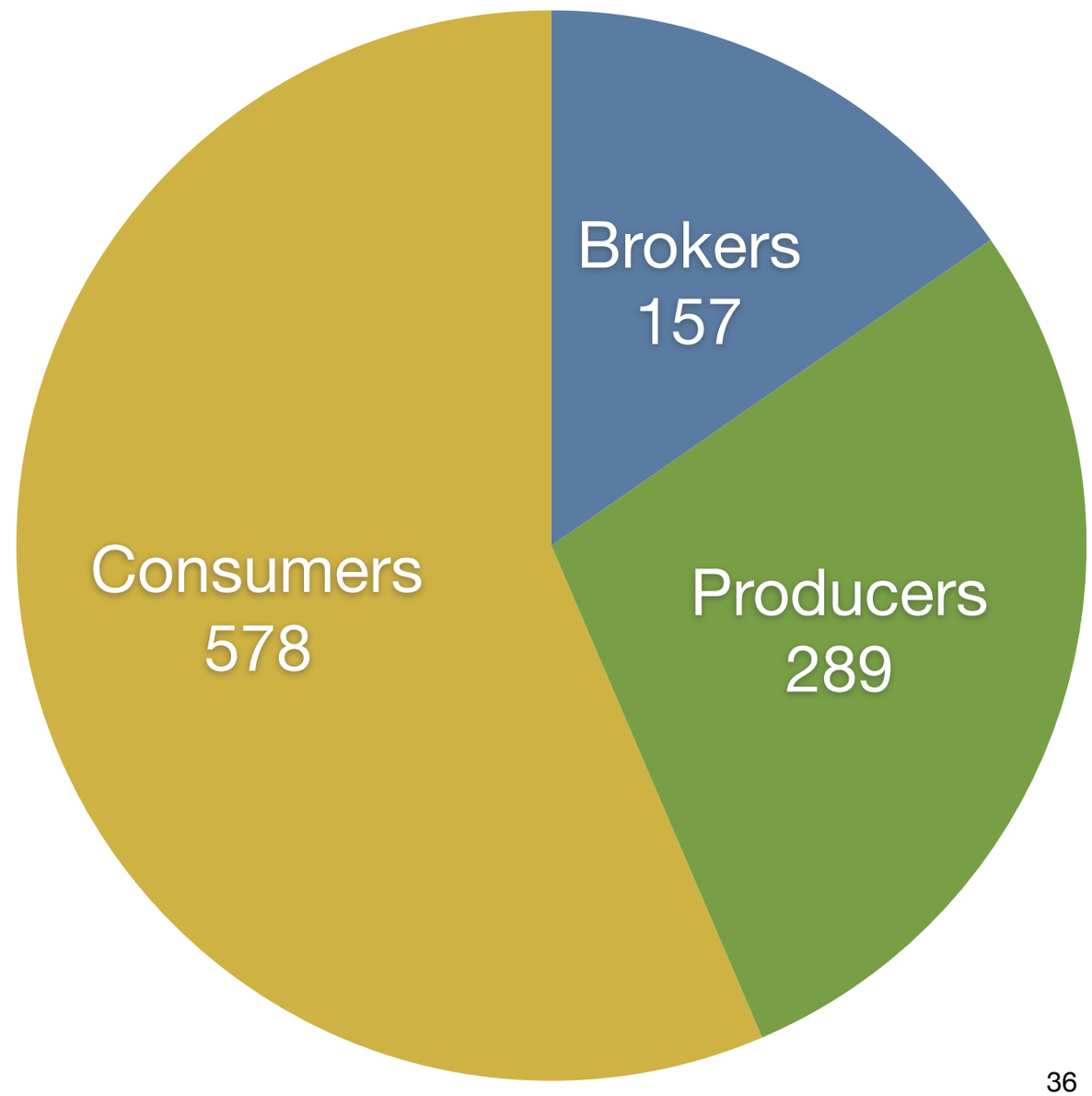
We need to allocate enough brokers to handle 32x speedup.



Edge Data Center Node Allocation

We need to allocate enough brokers to handle 32x speedup.





Targeted Data Center Design
Optimizing for Total Cost of Ownership

Targeted Data Center Design
Optimizing for Total Cost of Ownership

Homogeneous

Heterogeneous

Targeted Data Center Design
Optimizing for Total Cost of Ownership

Homogeneous

56 core
1 Drive
100 GbE

Heterogeneous

Targeted Data Center Design
Optimizing for Total Cost of Ownership

Homogeneous

56 core
1 Drive
100 GbE

Heterogeneous

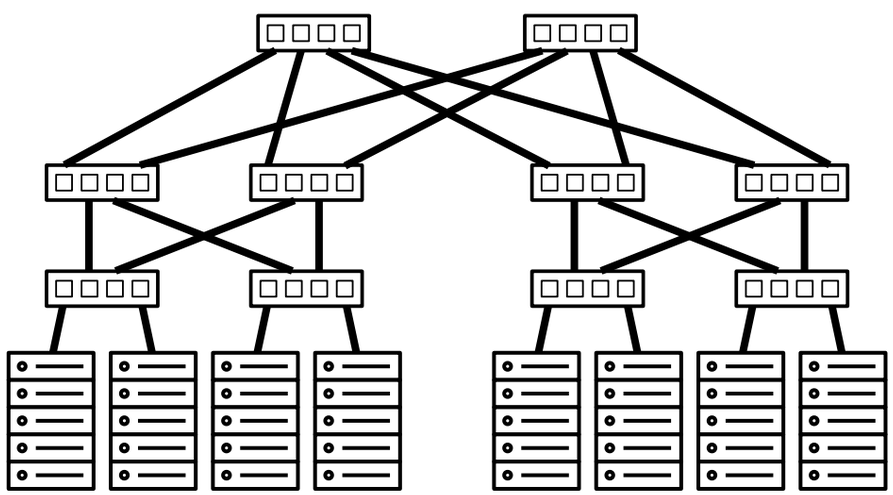
Compute

Broker

Targeted Data Center Design
Optimizing for Total Cost of Ownership

Homogeneous

56 core
1 Drive
100 GbE



160 switches

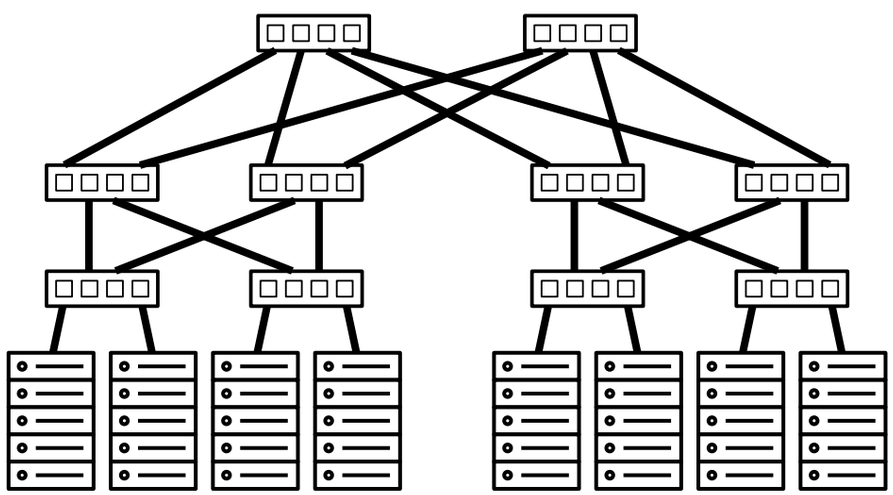
Heterogeneous

Compute Broker

Targeted Data Center Design
Optimizing for Total Cost of Ownership

Homogeneous

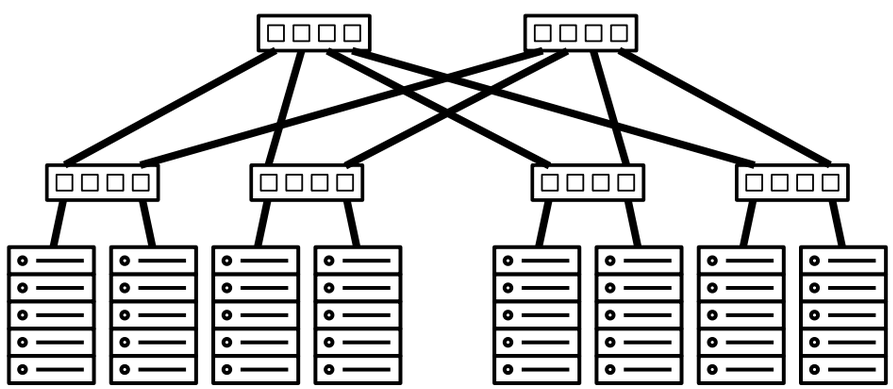
56 core
1 Drive
100 GbE



160 switches

Heterogeneous

Compute Broker

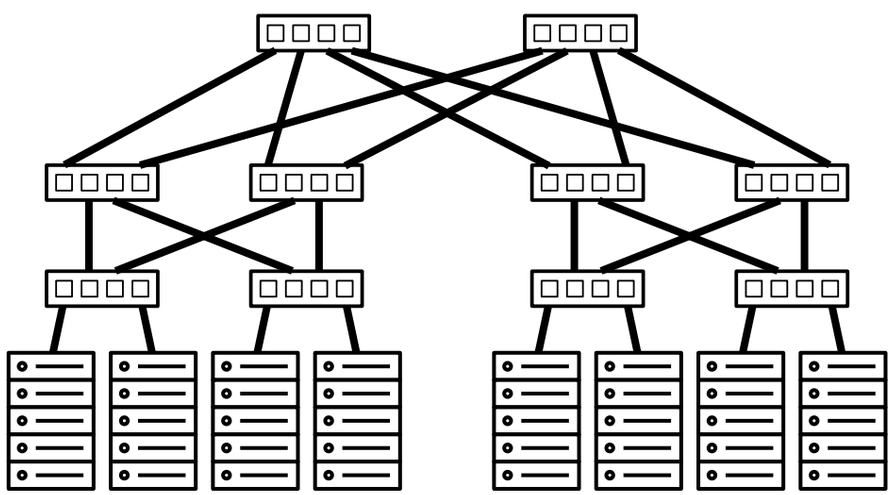


28+14 switches

Targeted Data Center Design
Optimizing for Total Cost of Ownership

Homogeneous

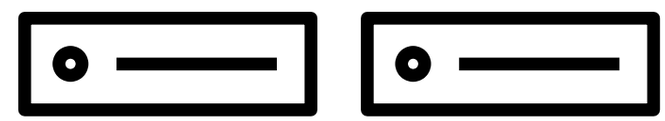
56 core
1 Drive
100 GbE

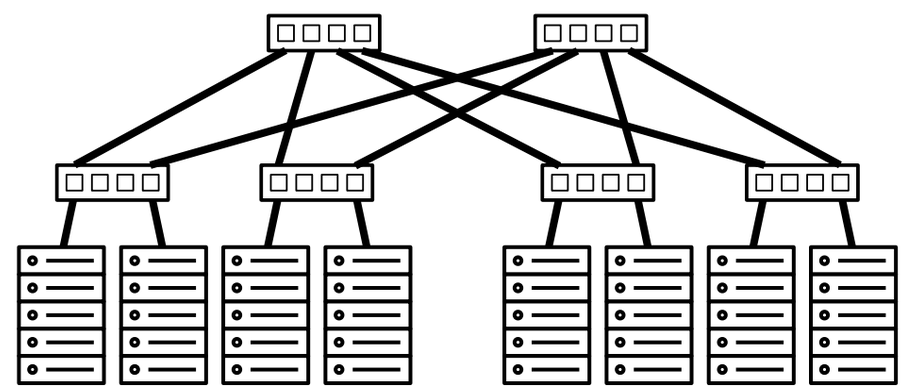


160 switches

Heterogeneous

Compute Broker





28+14 switches

Heterogeneous Edge Data Center

Heterogeneous Edge Data Center

Compute Node

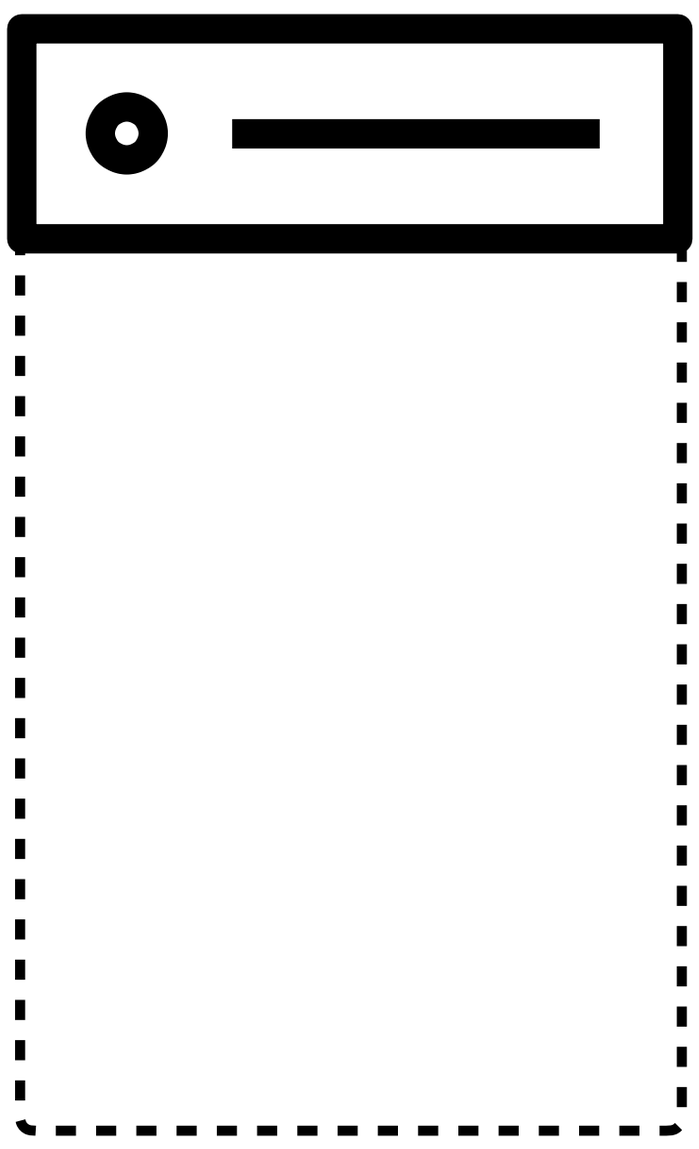
Heterogeneous Edge Data Center

Compute Node

Broker Node

Heterogeneous Edge Data Center

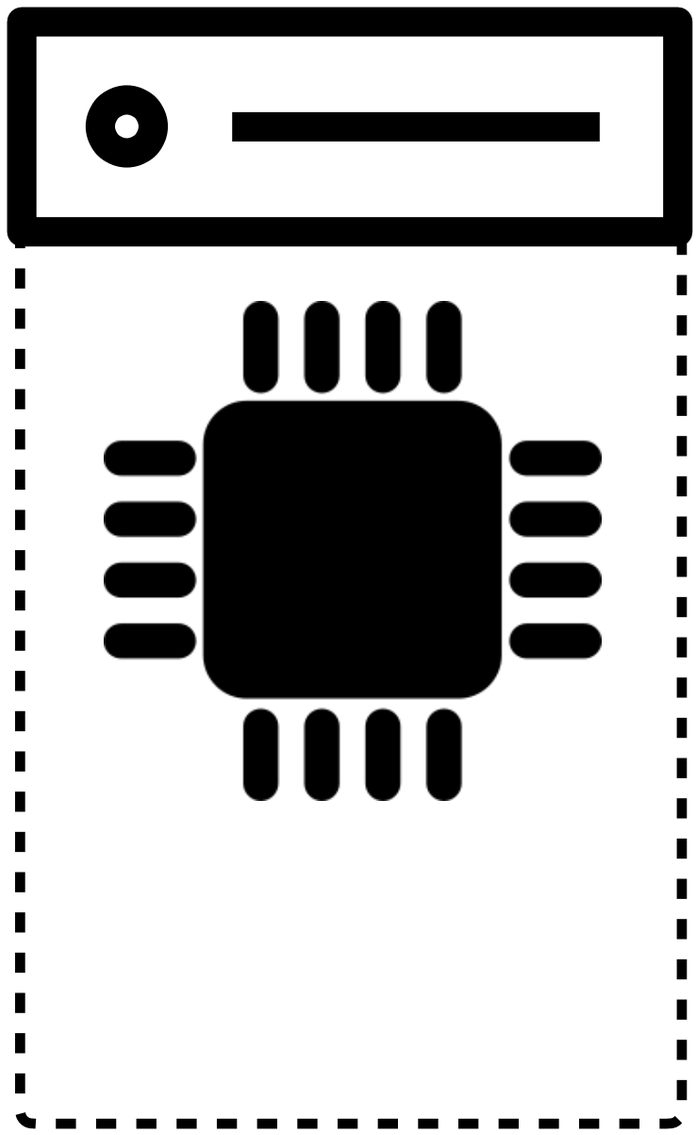
Compute Node



Broker Node

Heterogeneous Edge Data Center

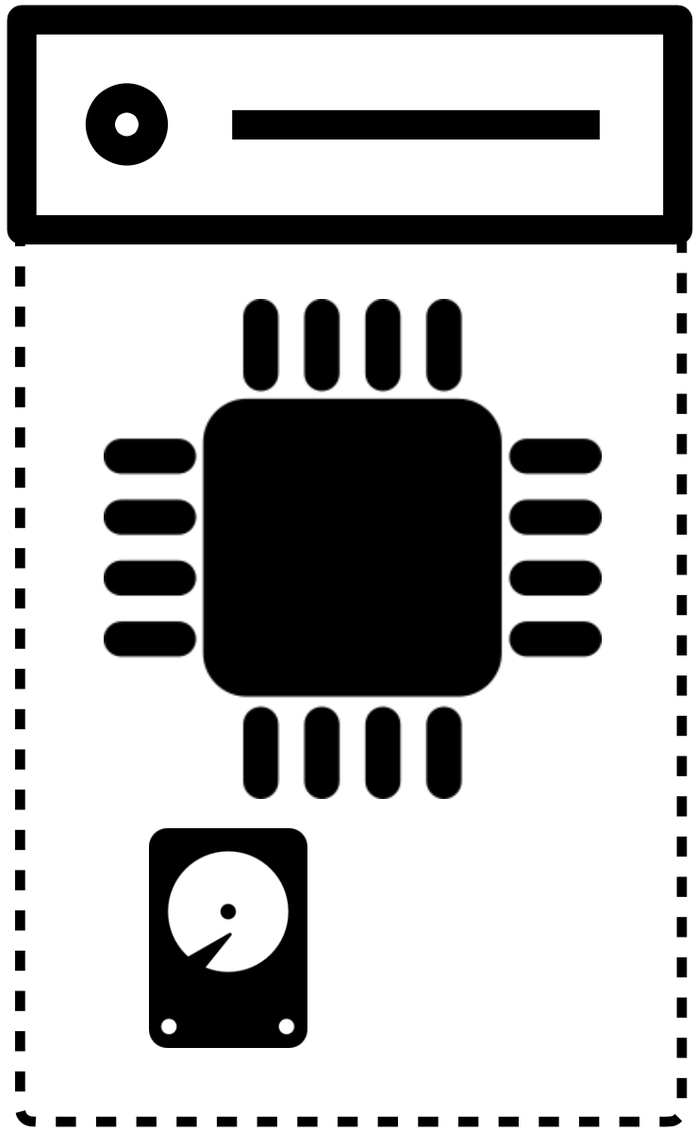
Compute Node



Broker Node

Heterogeneous Edge Data Center

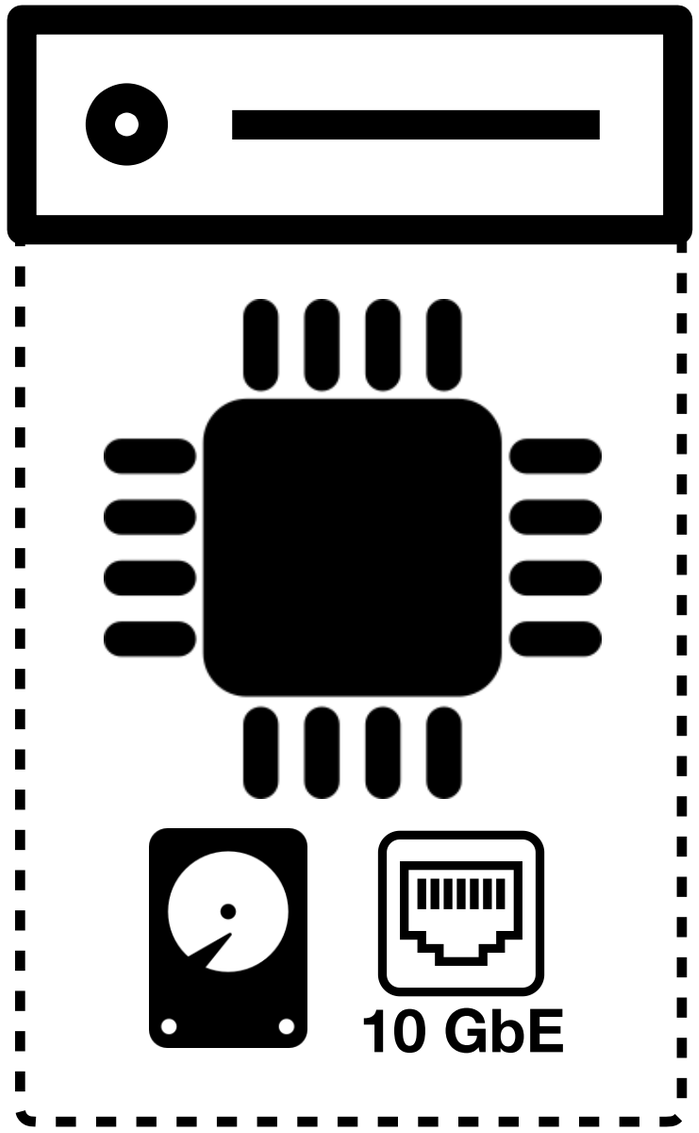
Compute Node



Broker Node

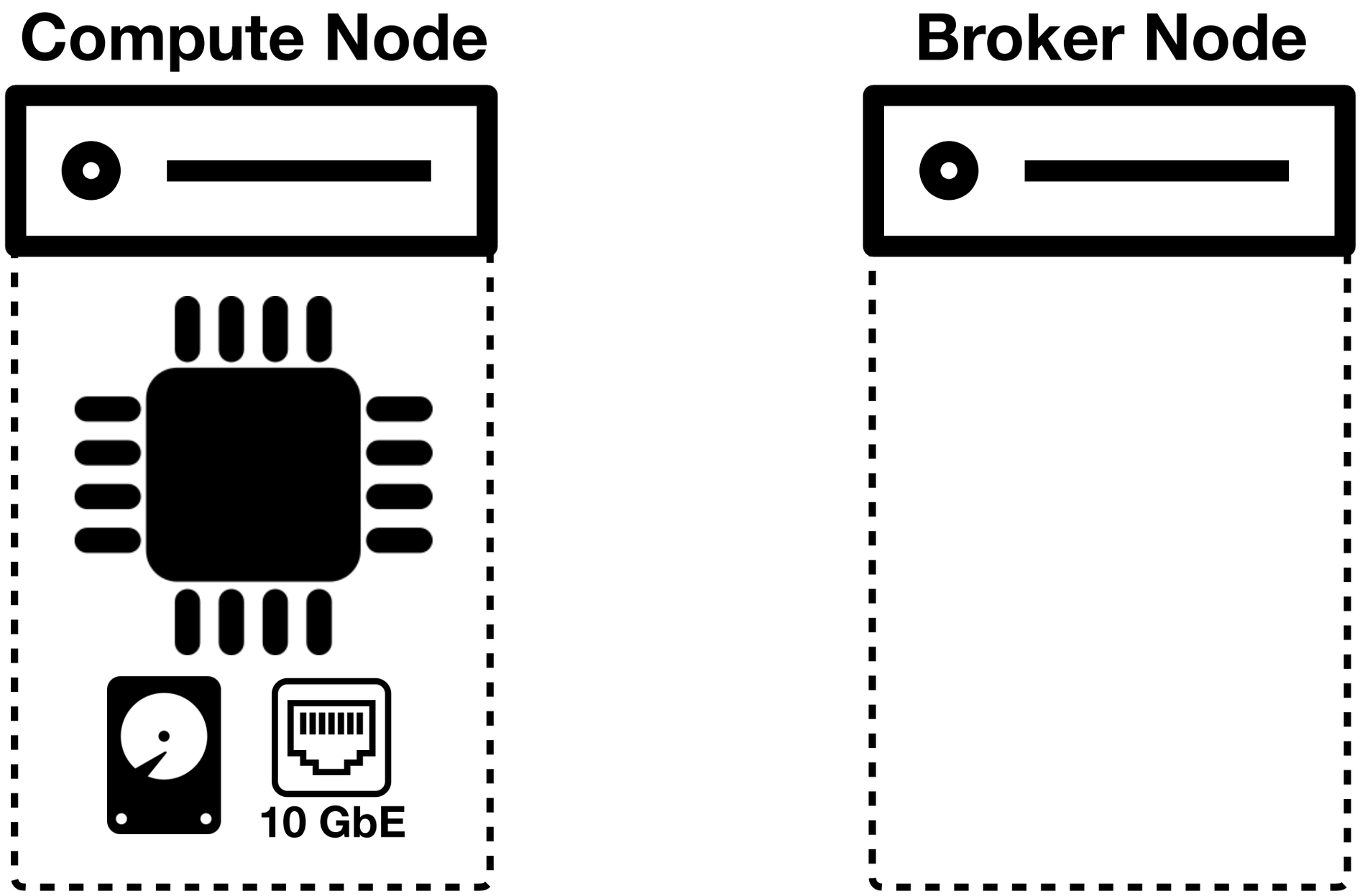
Heterogeneous Edge Data Center

Compute Node

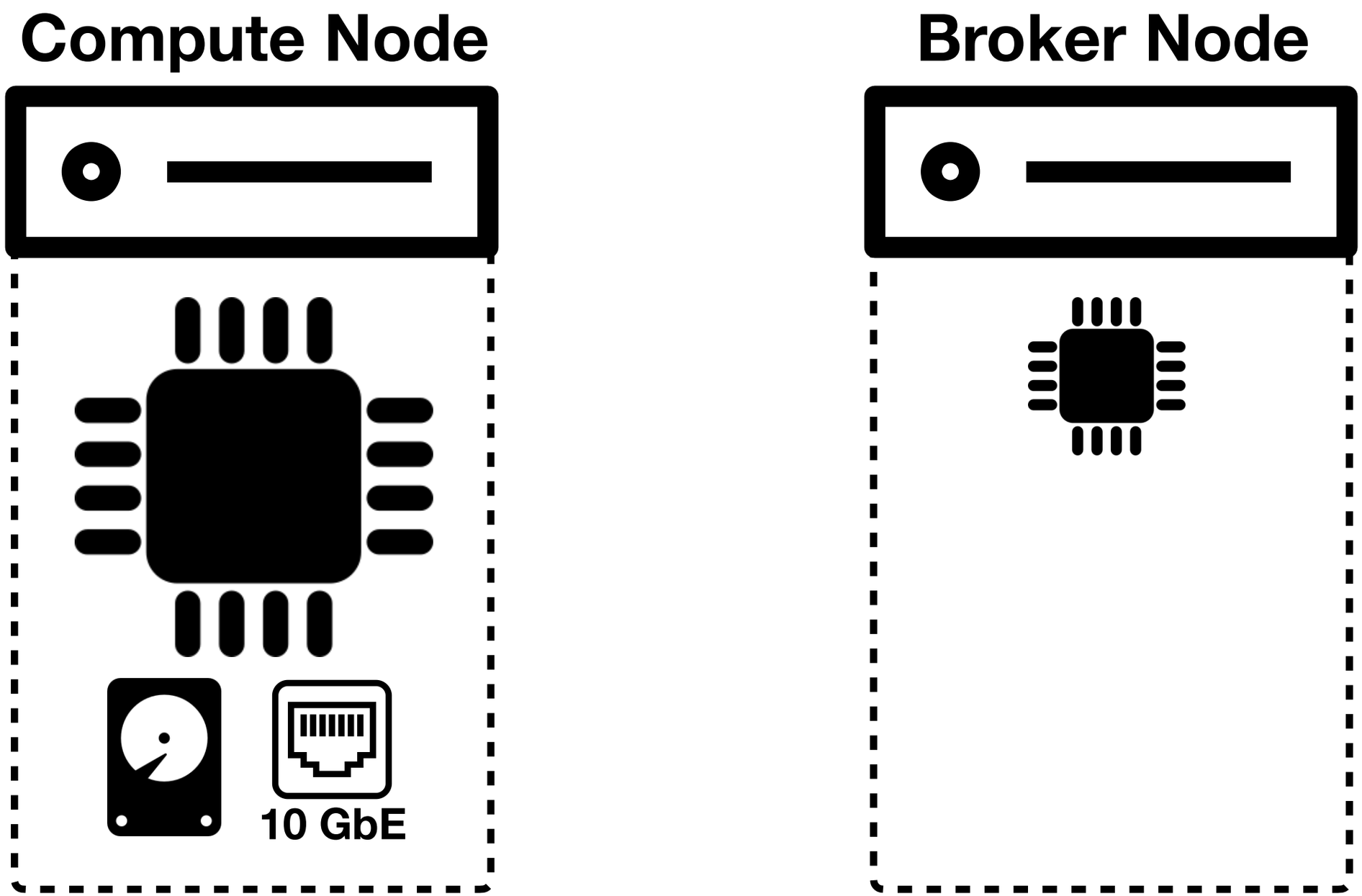


Broker Node

Heterogeneous Edge Data Center

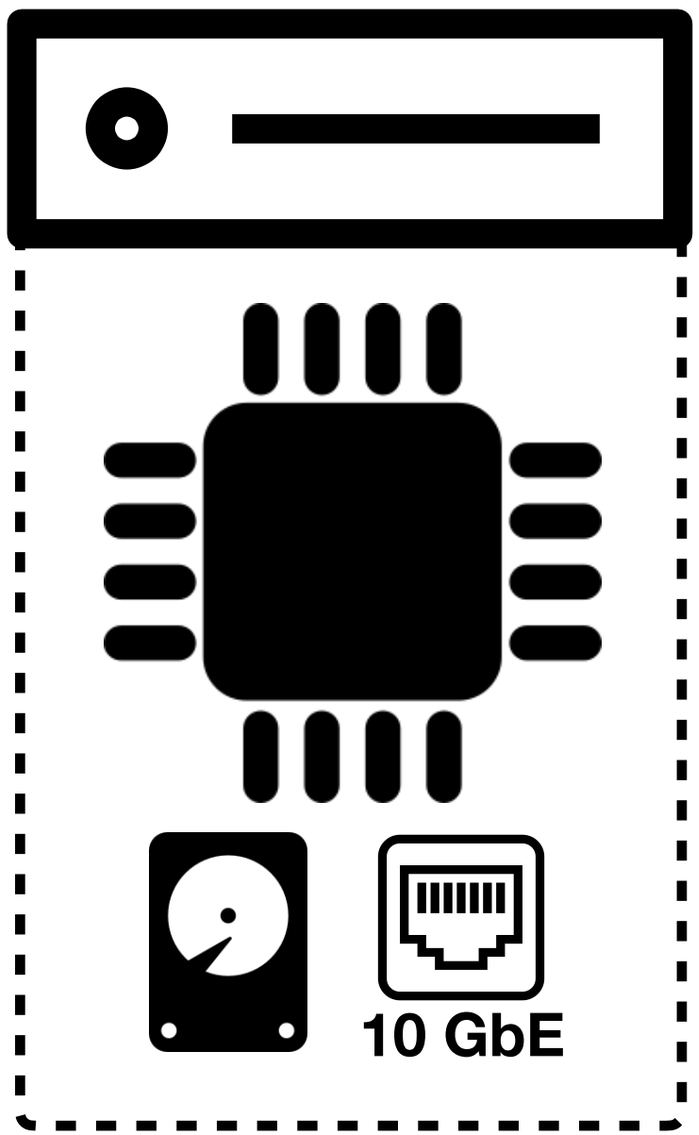


Heterogeneous Edge Data Center

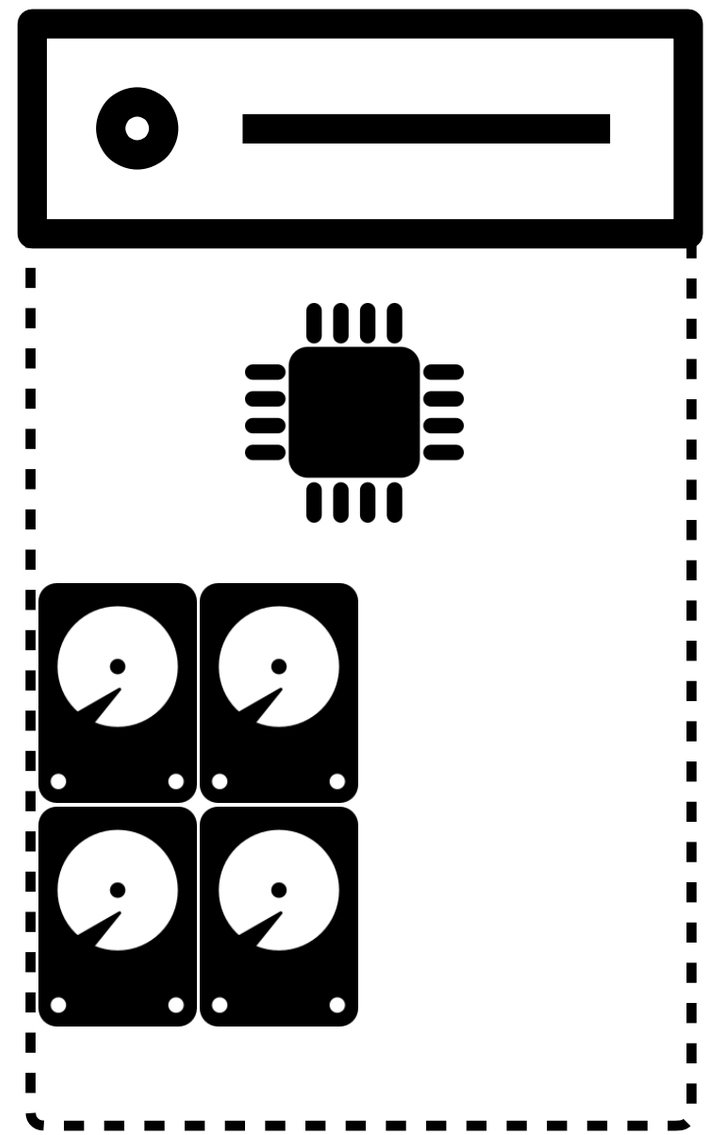


Heterogeneous Edge Data Center

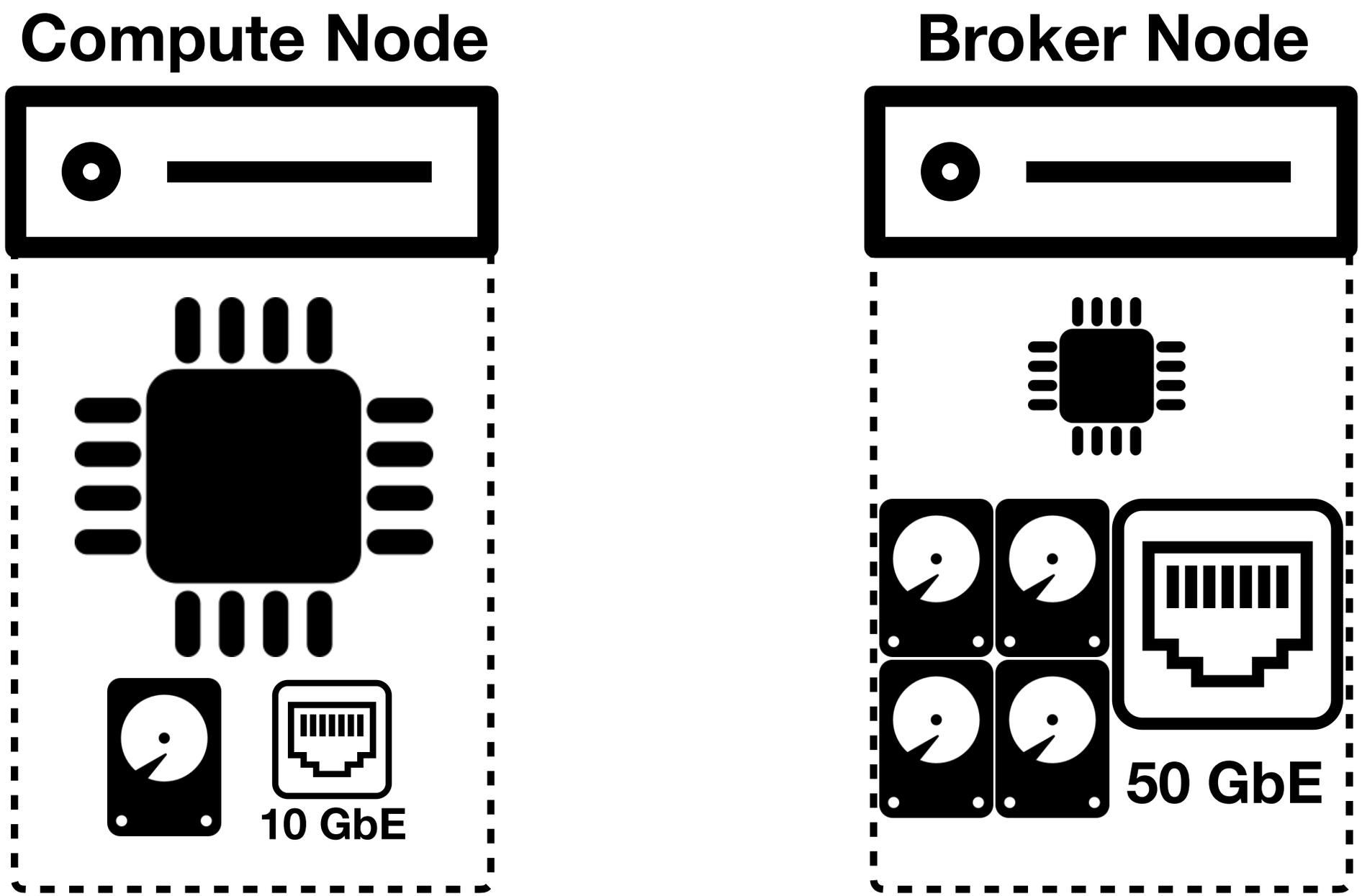
Compute Node



Broker Node

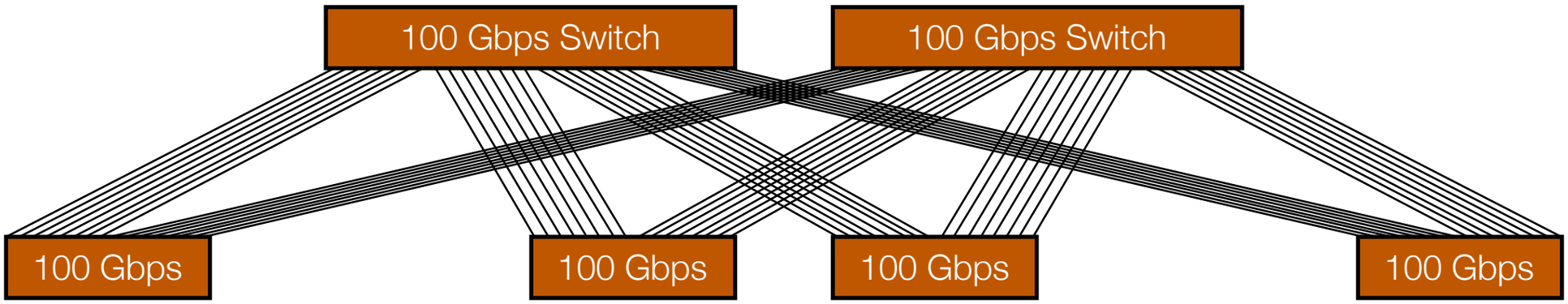


Heterogeneous Edge Data Center

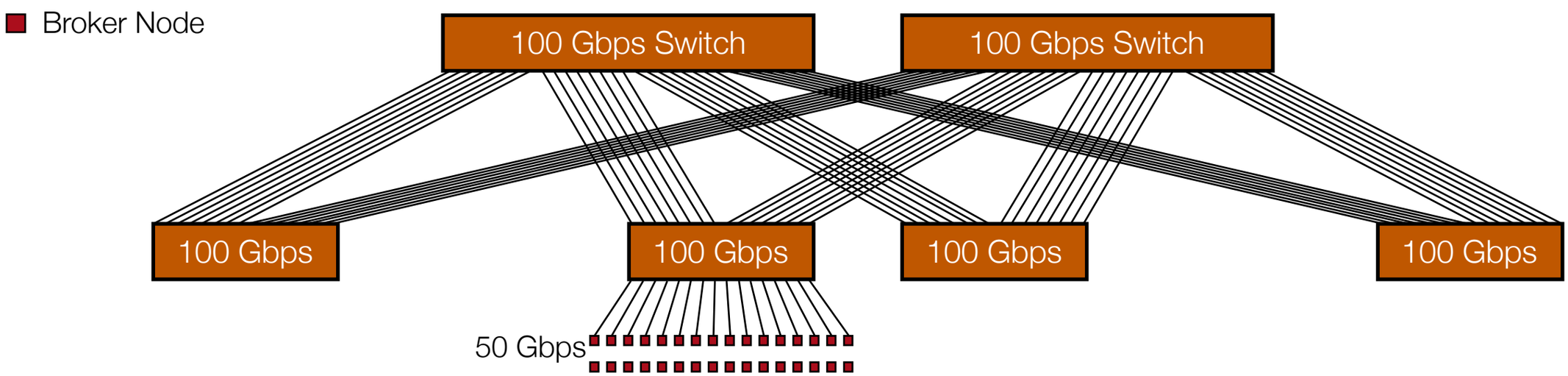


Heterogeneous Edge Data Center Networking

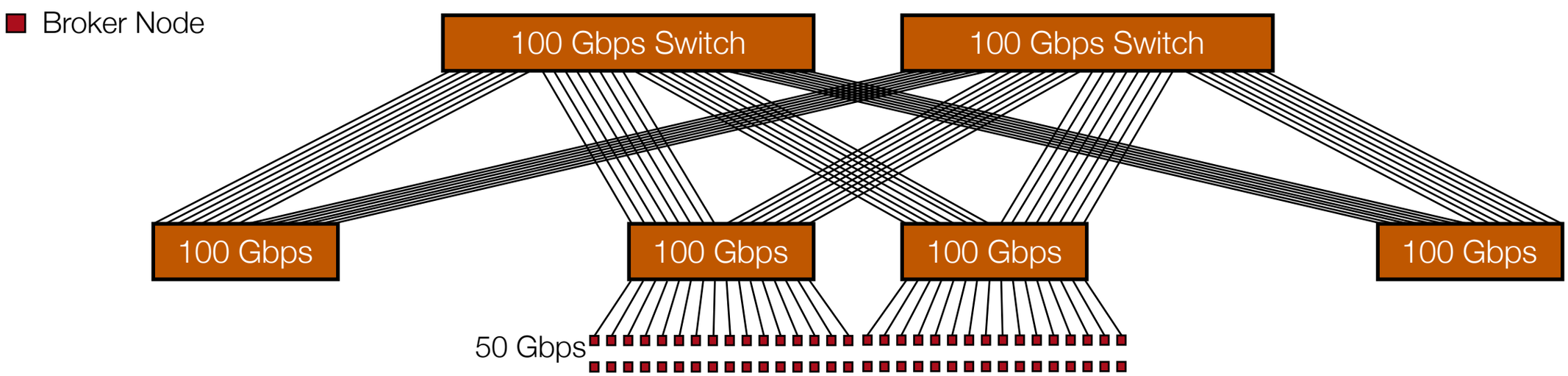
Heterogeneous Edge Data Center Networking



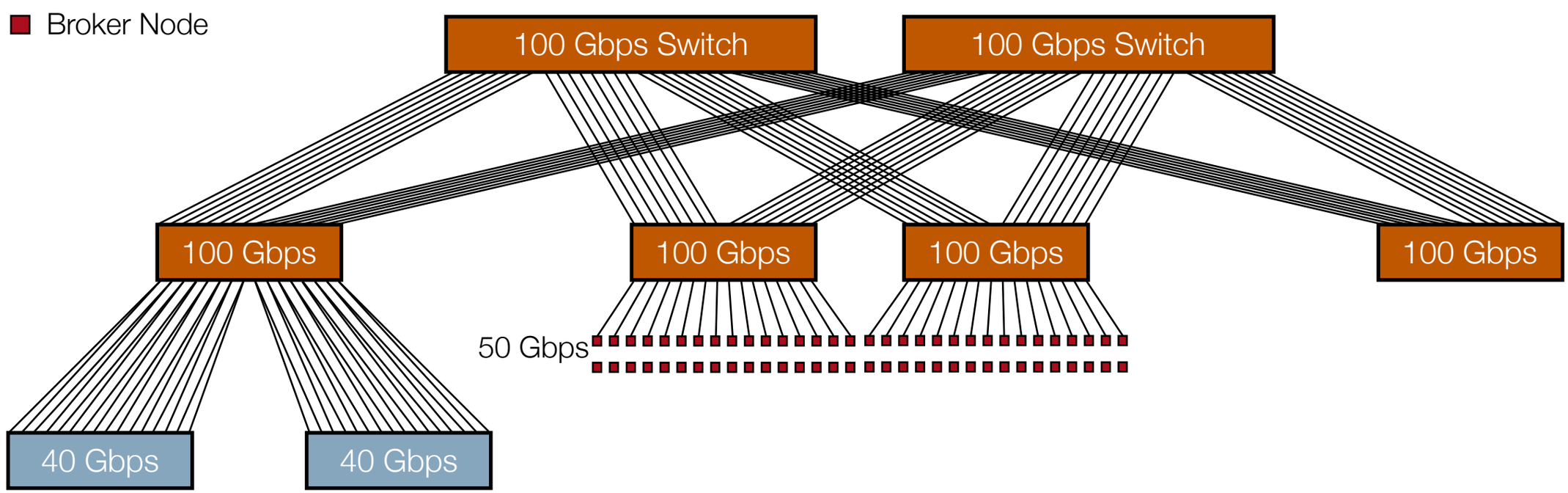
Heterogeneous Edge Data Center Networking



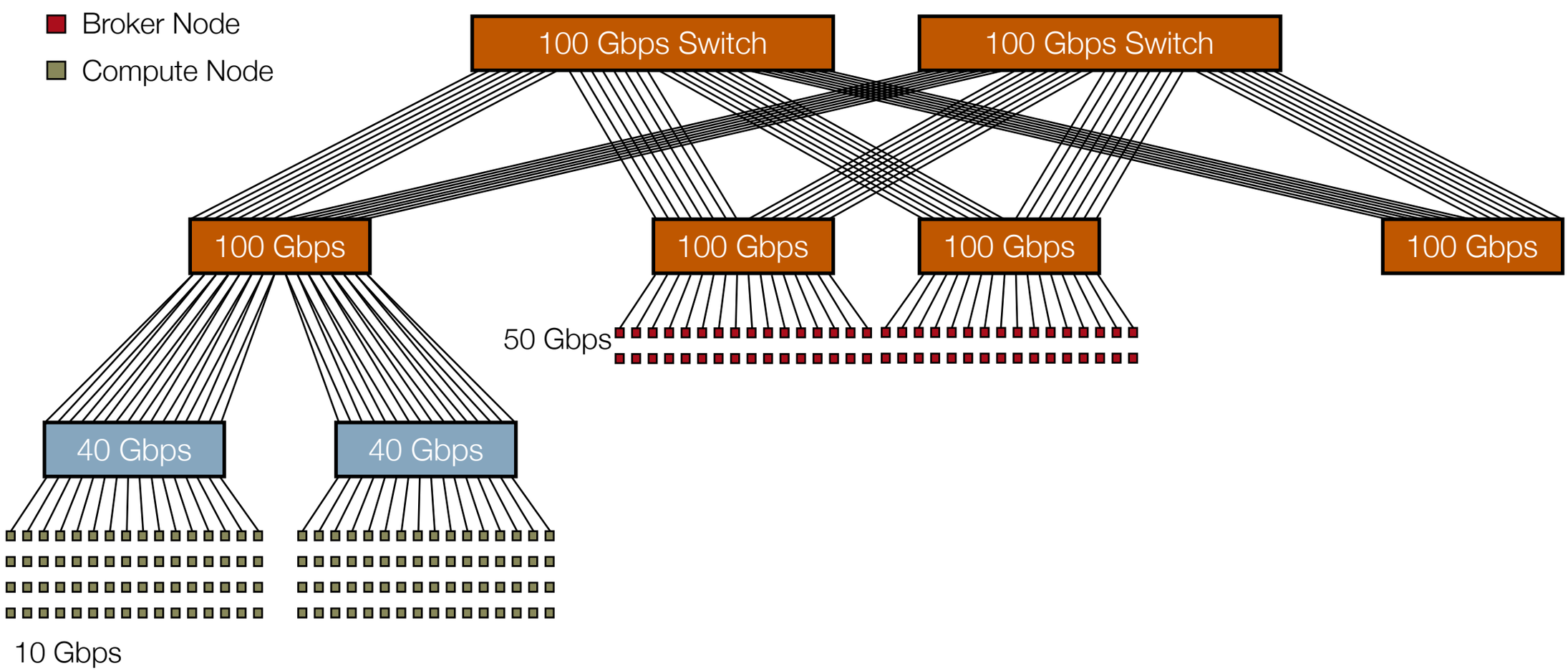
Heterogeneous Edge Data Center Networking



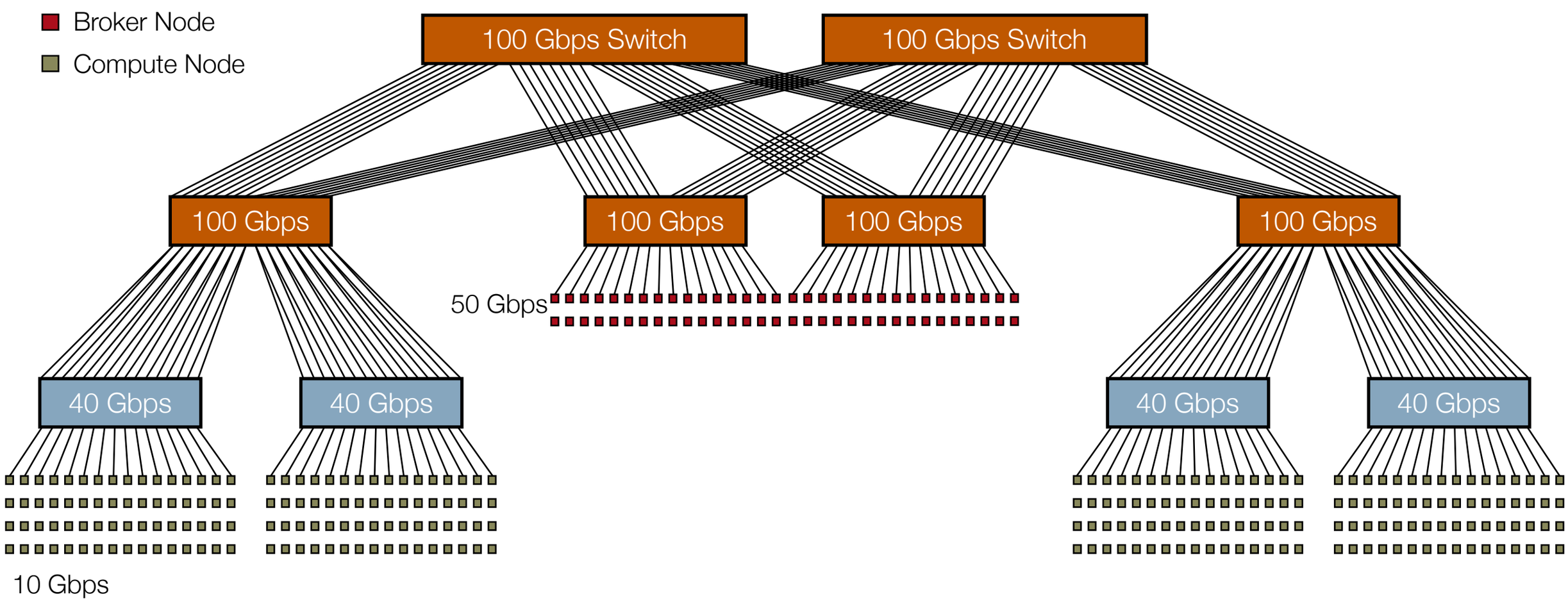
Heterogeneous Edge Data Center Networking



Heterogeneous Edge Data Center Networking



Heterogeneous Edge Data Center Networking

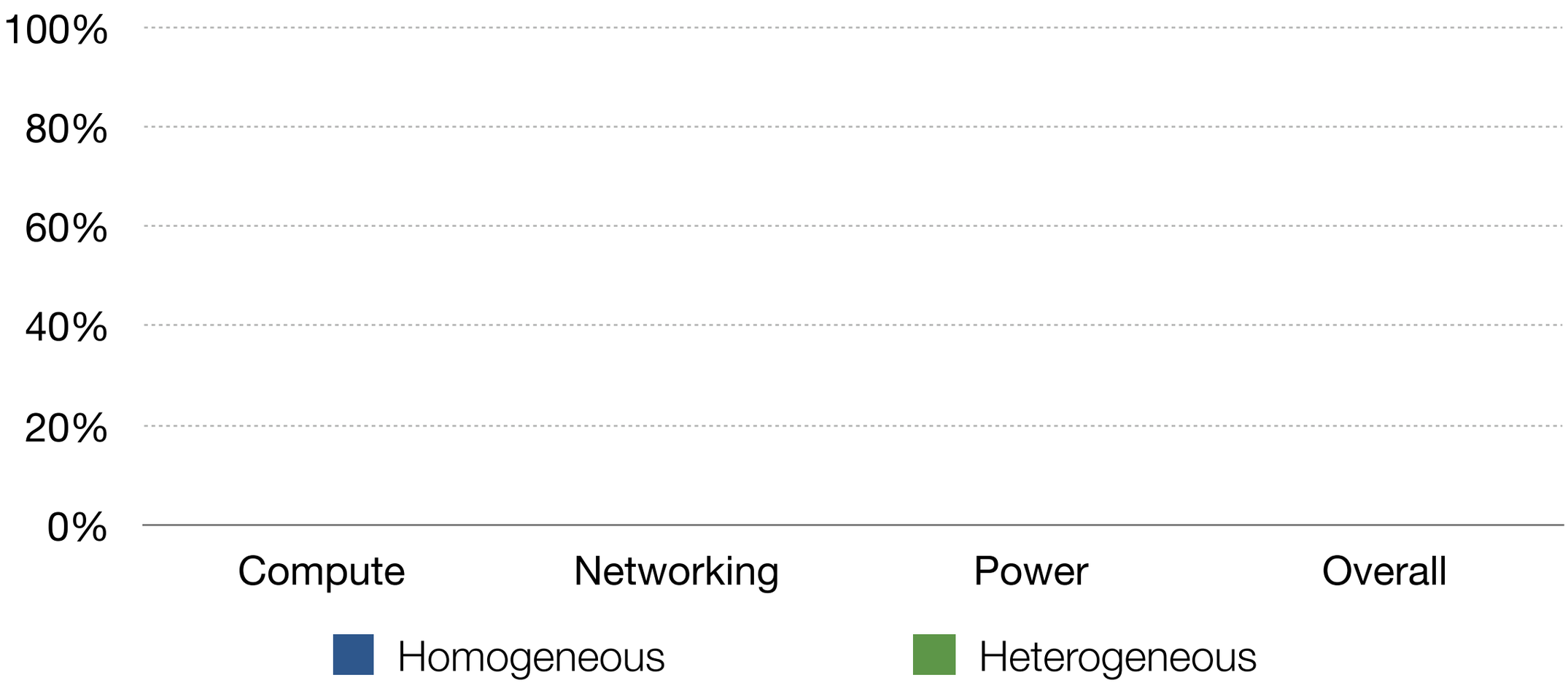


Comparing Total Cost of Ownership

We assume a three-year amortization of costs.

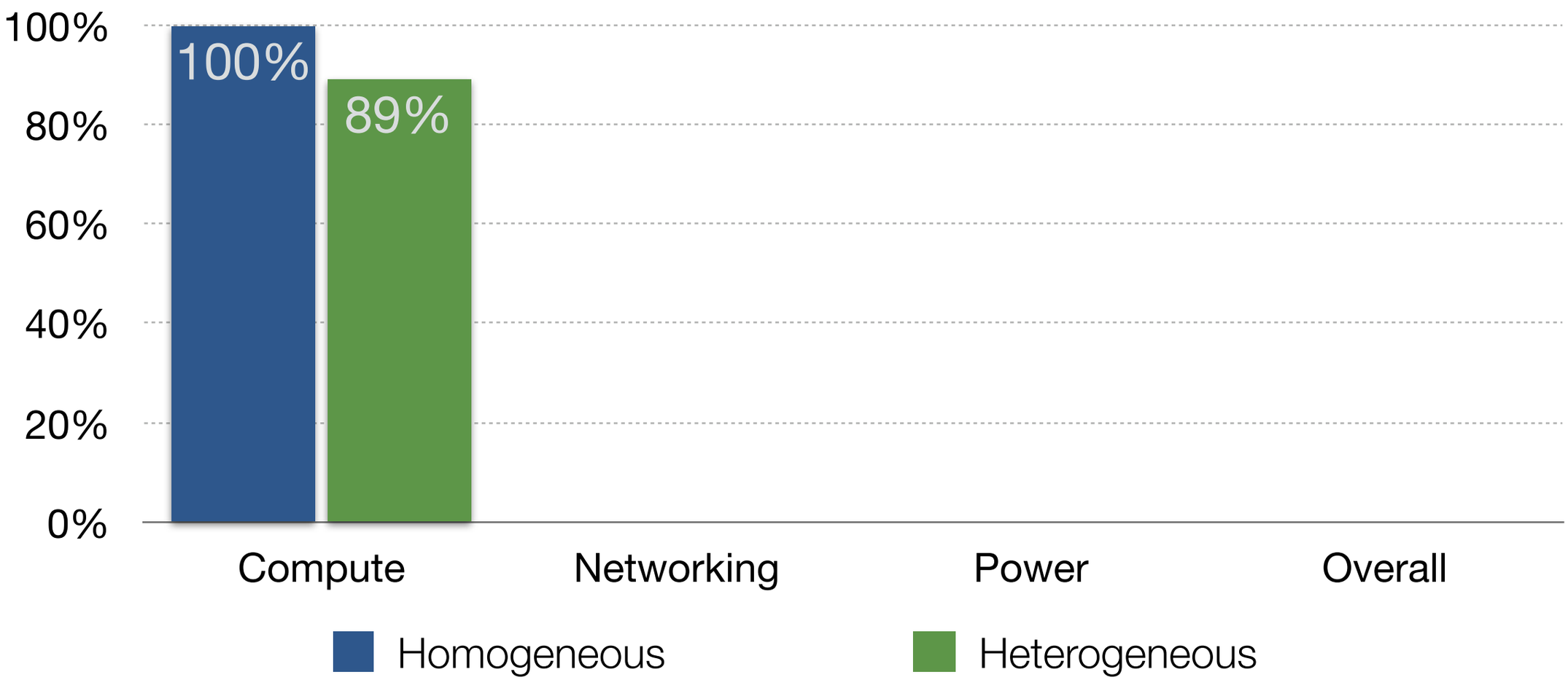
Comparing Total Cost of Ownership

We assume a three-year amortization of costs.



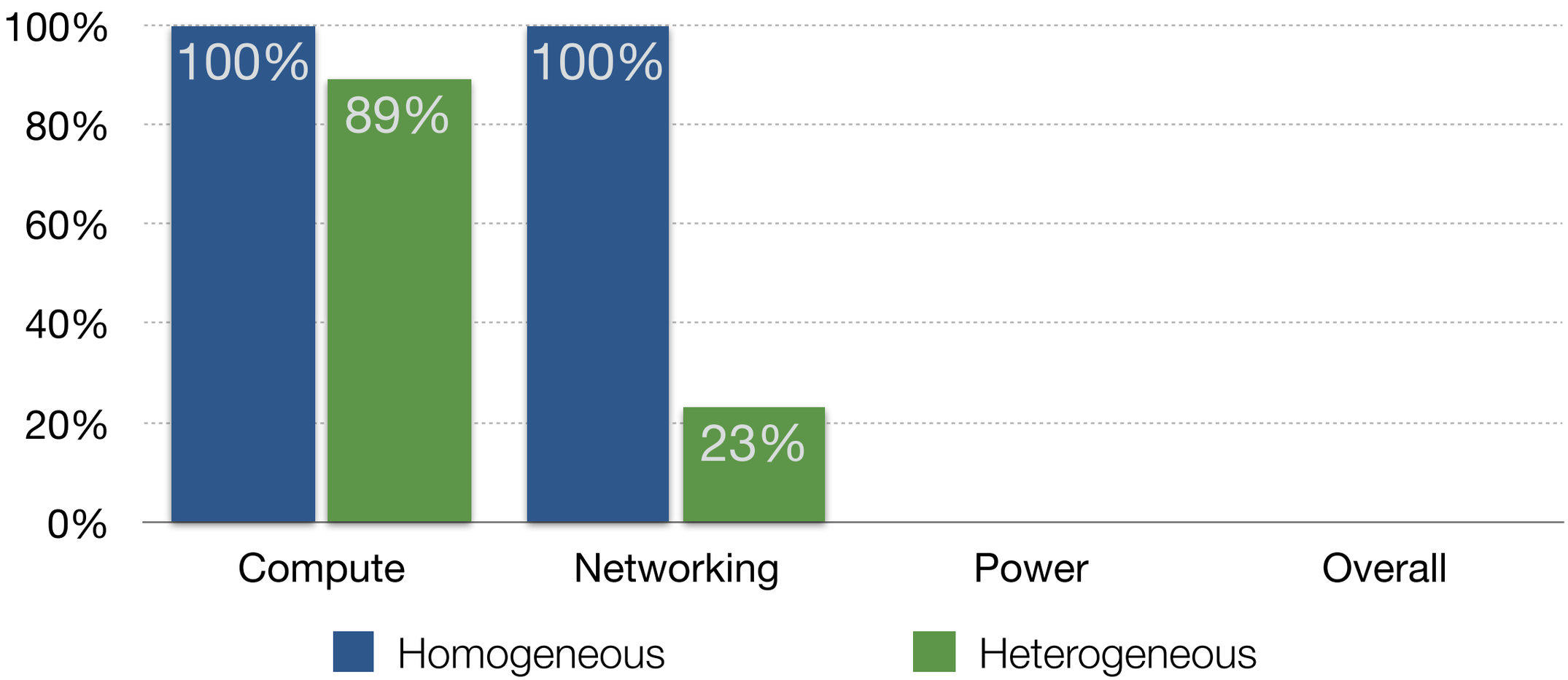
Comparing Total Cost of Ownership

We assume a three-year amortization of costs.



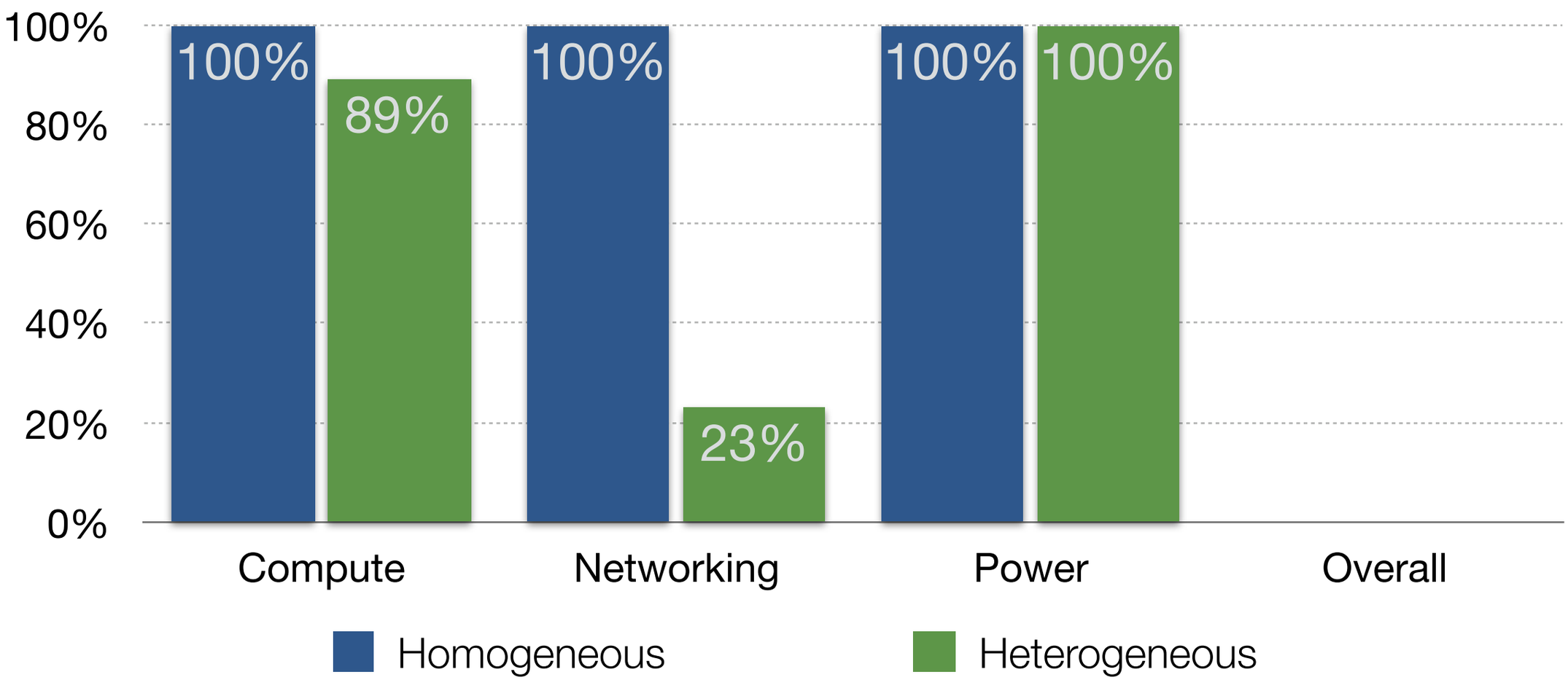
Comparing Total Cost of Ownership

We assume a three-year amortization of costs.



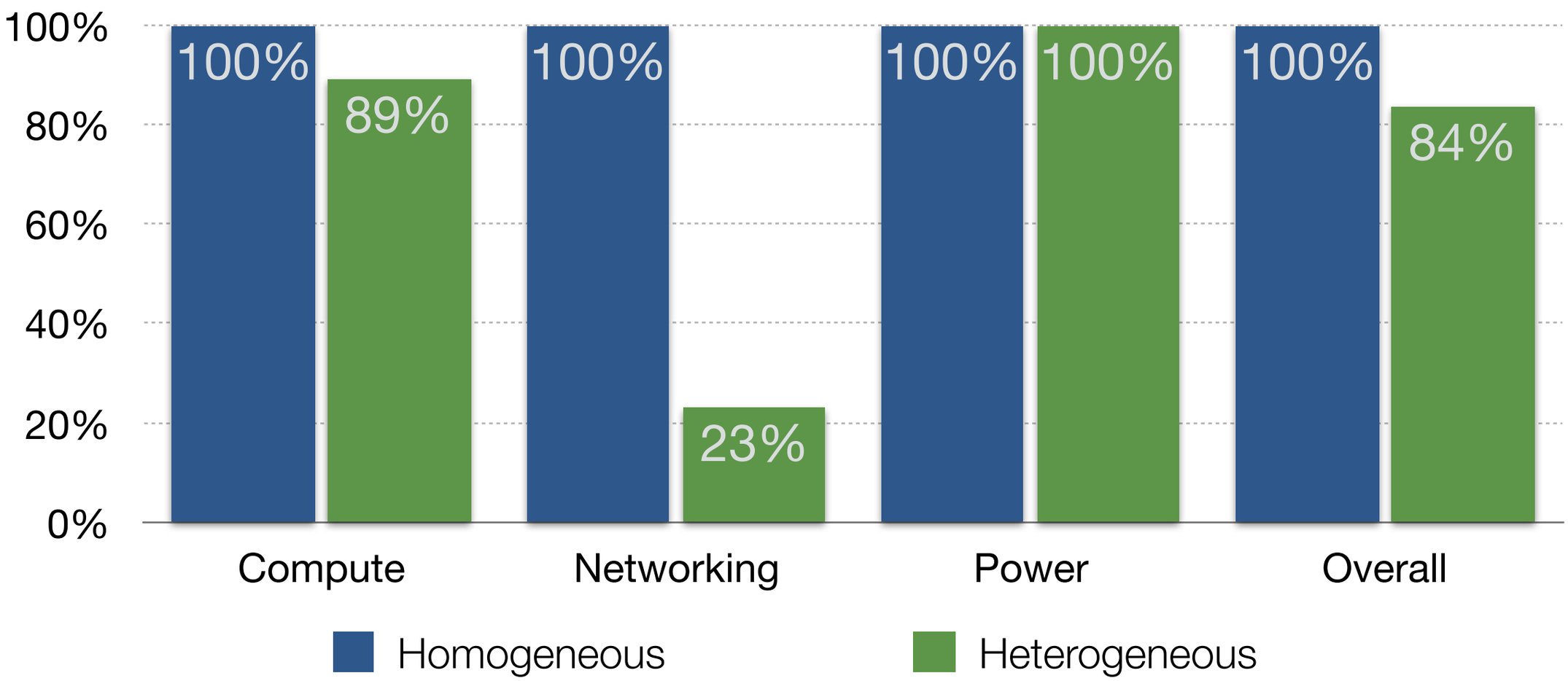
Comparing Total Cost of Ownership

We assume a three-year amortization of costs.



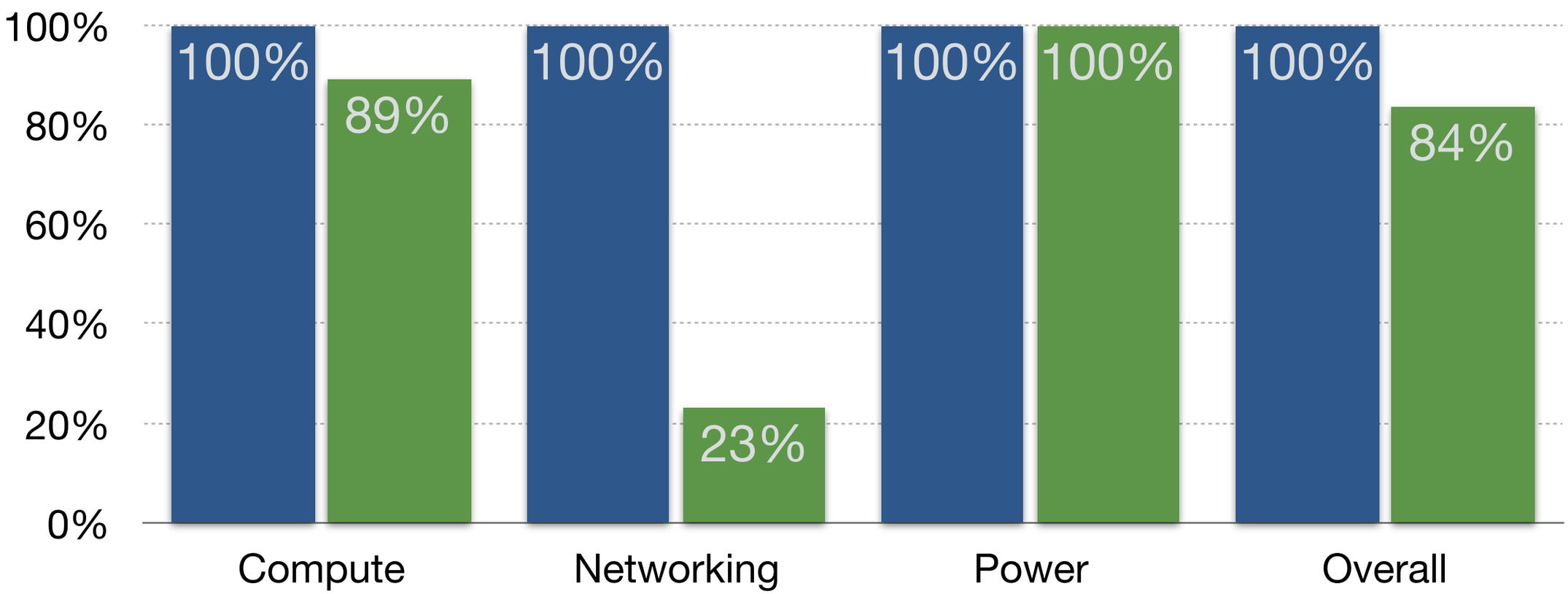
Comparing Total Cost of Ownership

We assume a three-year amortization of costs.



Comparing Total Cost of Ownership

We assume a three-year amortization of costs.



Homogeneous

$12.9 million

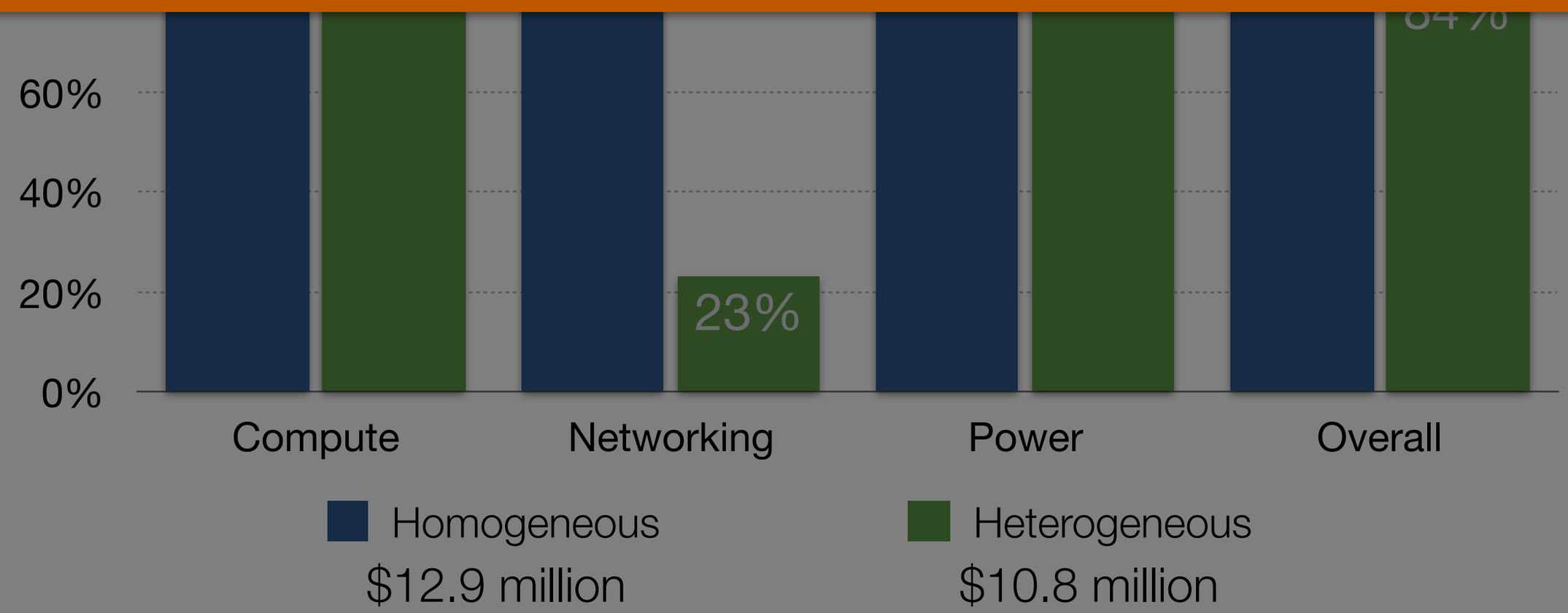
Heterogeneous

$10.8 million

Comparing Total Cost of Ownership

We assume a three-year amortization of costs.

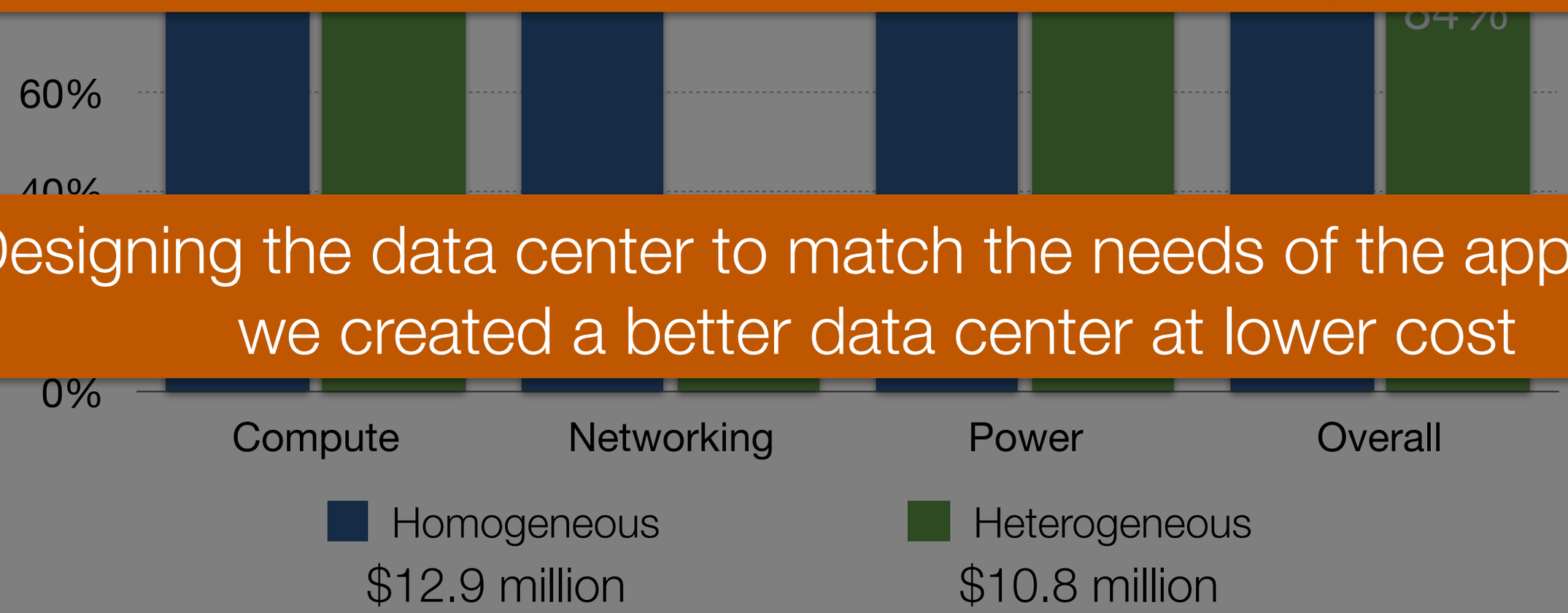
The targeted, heterogeneous data center incurs
16% lower total cost of ownership



Comparing Total Cost of Ownership

We assume a three-year amortization of costs.

The targeted, heterogeneous data center incurs
16% lower total cost of ownership



Designing the data center to match the needs of the application,
we created a better data center at lower cost

Conclusion

Calls to Action

Calls to Action

- To fully understand AI applications, **we must consider the overhead of the AI tax** in end-to-end performance.

Calls to Action

- To fully understand AI applications, **we must consider the overhead of the AI tax** in end-to-end performance.
- As we accelerate AI, **we must consider new bottlenecks** that manifest as AI tax.

Calls to Action

- To fully understand AI applications, **we must consider the overhead of the AI tax** in end-to-end performance.
- As we accelerate AI, **we must consider new bottlenecks** that manifest as AI tax.
- We cannot limit our view of AI to microarchitectural considerations. **We need data center-level optimizations to address data center-level bottlenecks.**

Thank You