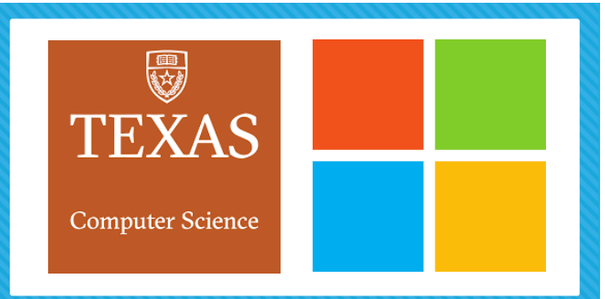


Rearchitecting the Datacenter Lifecycle for AI

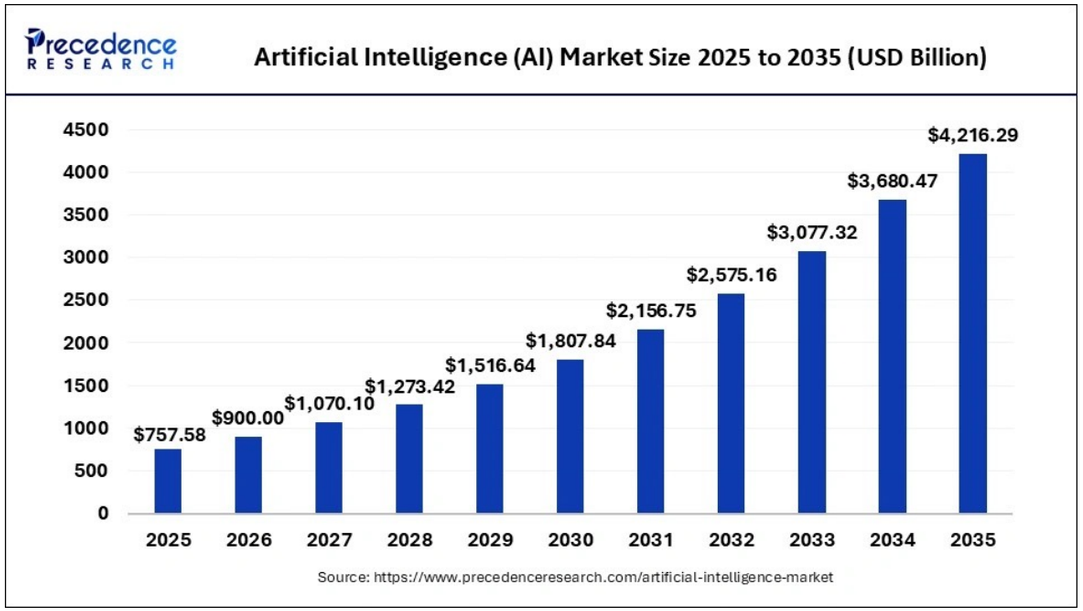
ISCA 2026

Jovan Stojkovic, Chaojie Zhang*, Íñigo Goiri*, Ricardo Bianchini*

The University of Texas at Austin, *Microsoft Azure

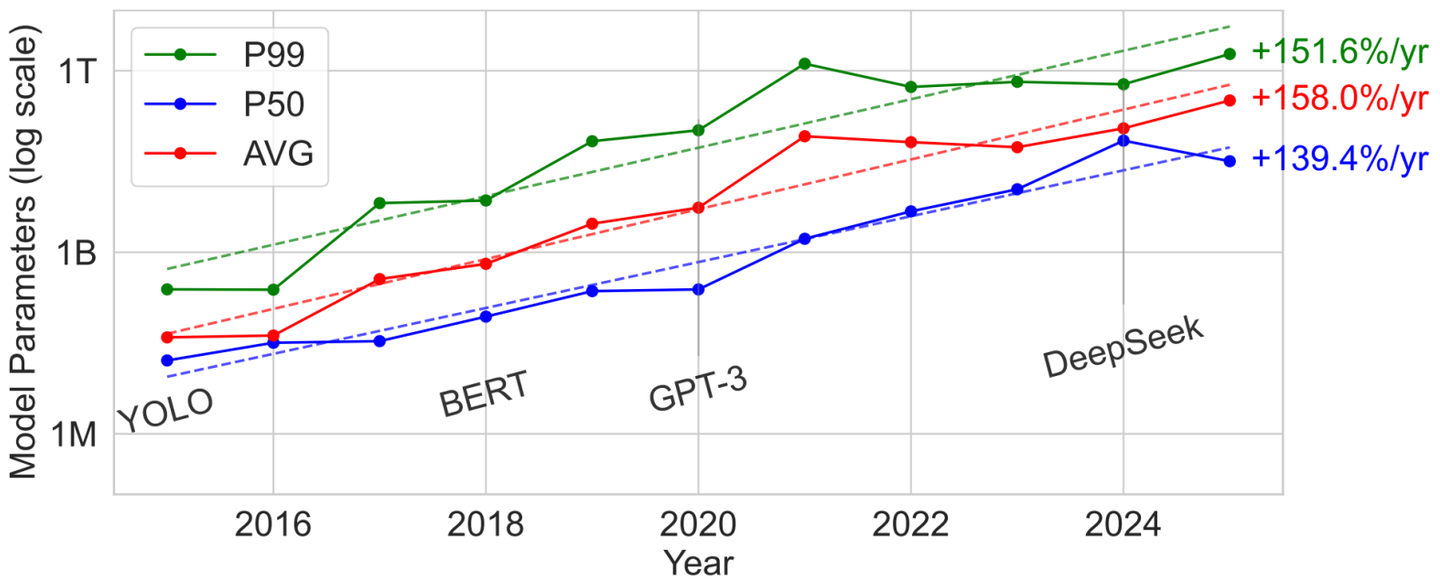
Motivation

- AI inference is becoming a dominant datacenter workload



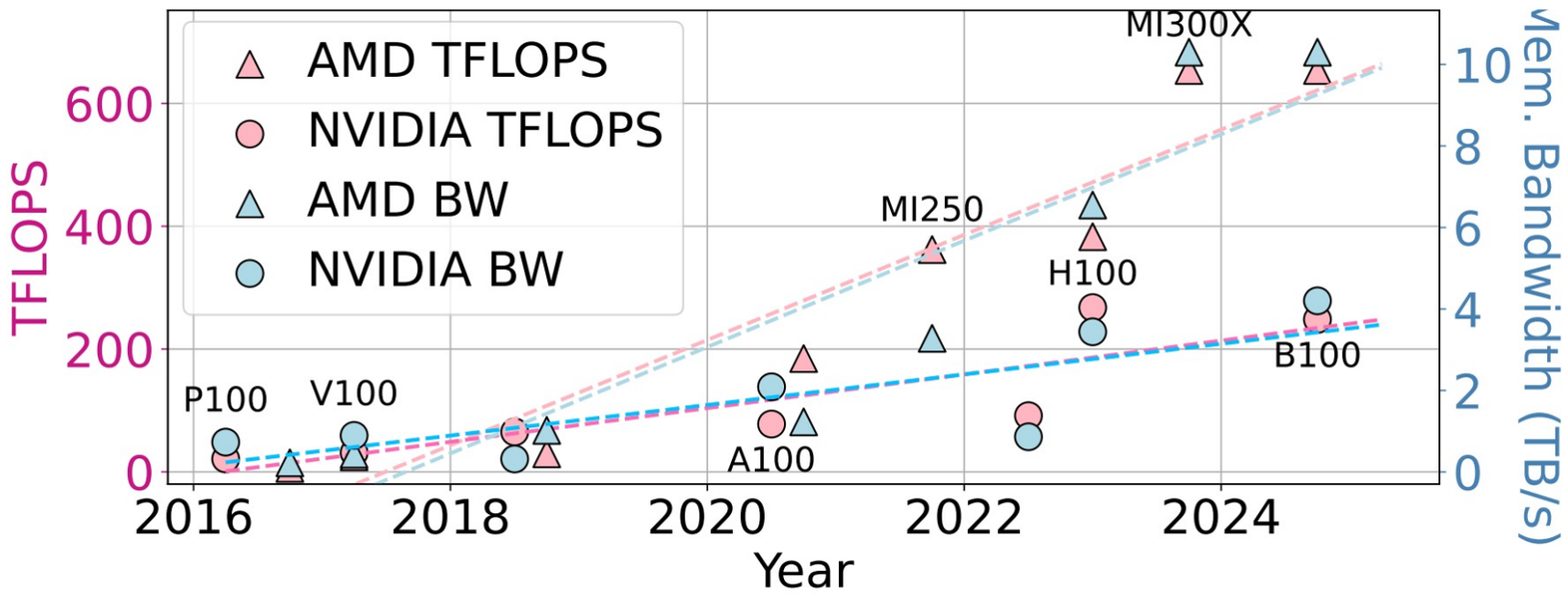
Motivation

- AI inference is becoming a dominant datacenter workload
- AI workloads and hardware are fast evolving



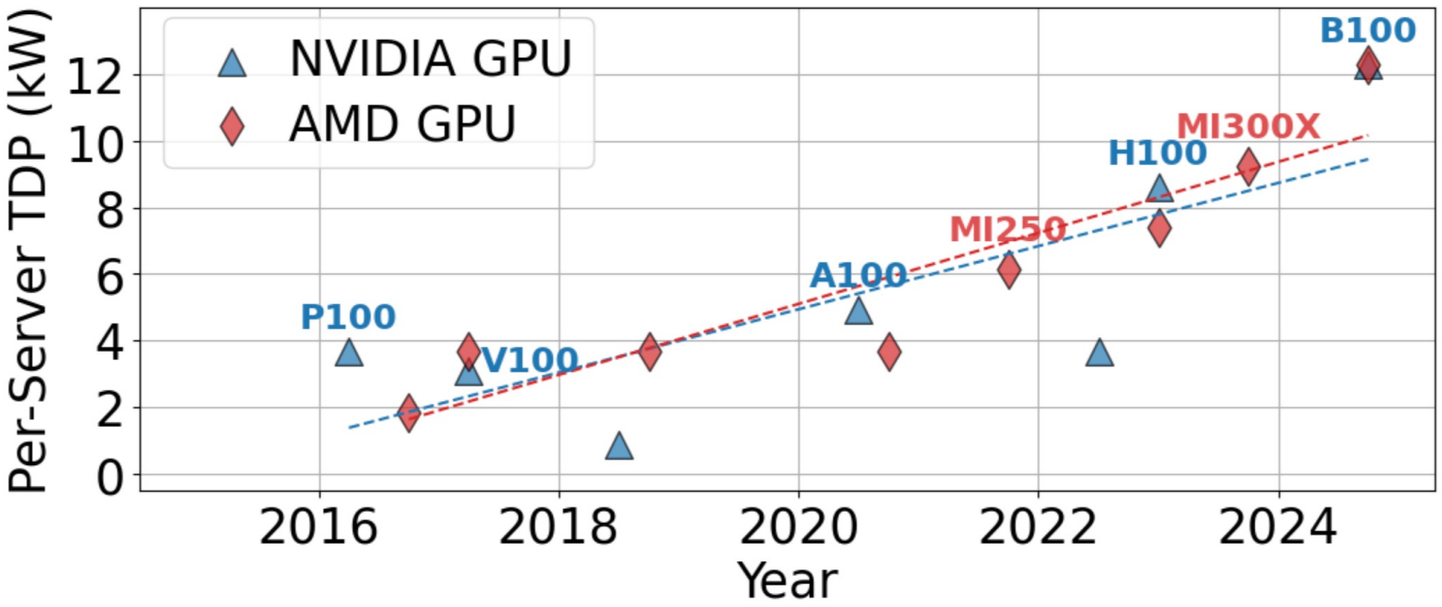
Motivation

- AI inference is becoming a dominant datacenter workload
- AI workloads and hardware are fast evolving



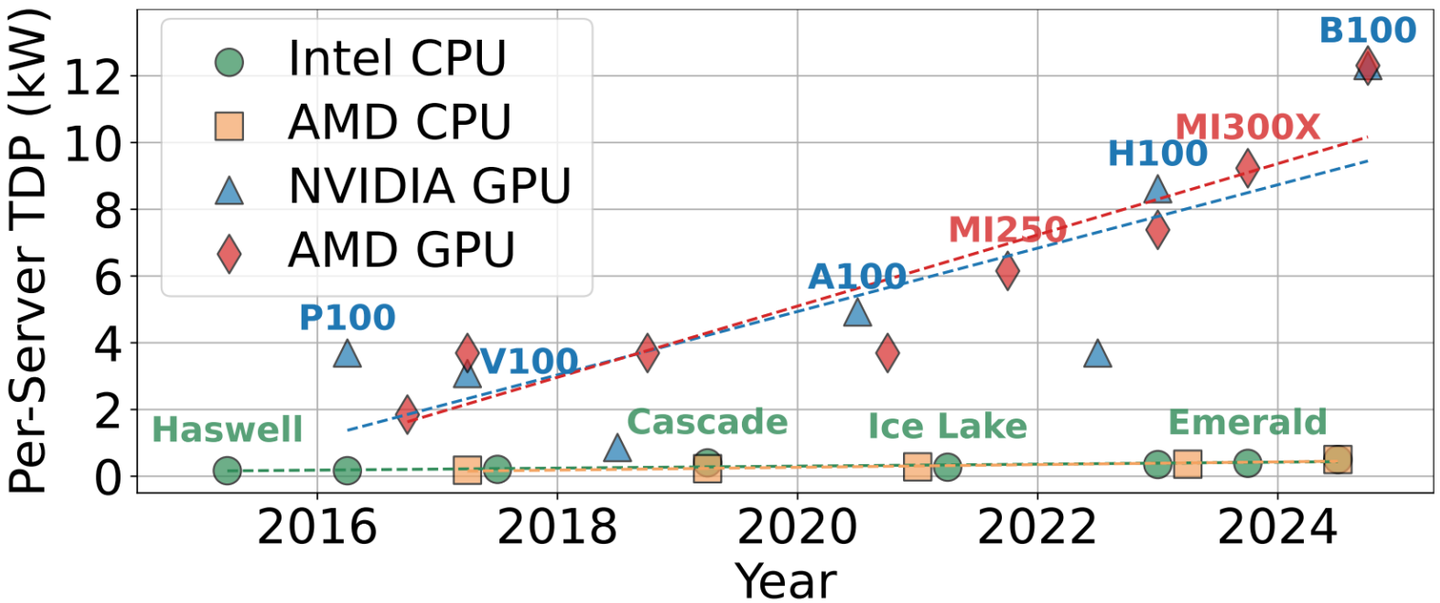
Motivation

- AI inference is becoming a dominant datacenter workload
- AI workloads and hardware are fast evolving



Motivation

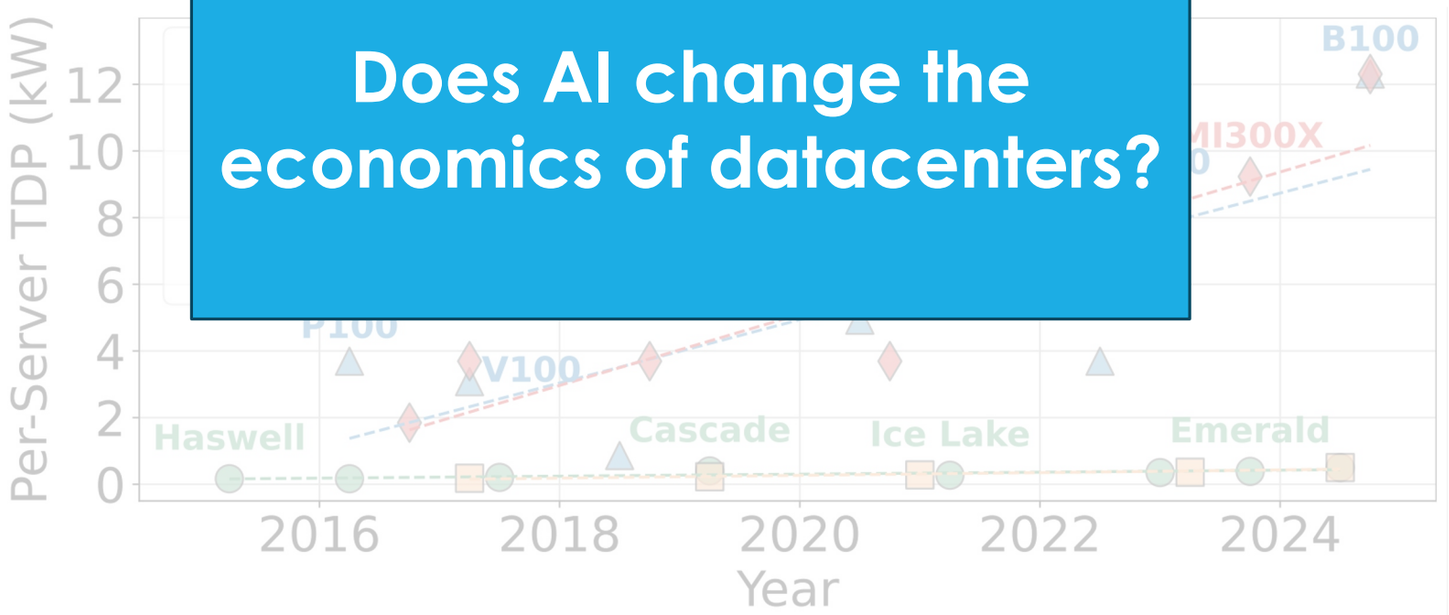
- AI inference is becoming a dominant datacenter workload
- AI workloads and hardware are fast evolving
- Traditional datacenter management was designed for CPUs



Motivation

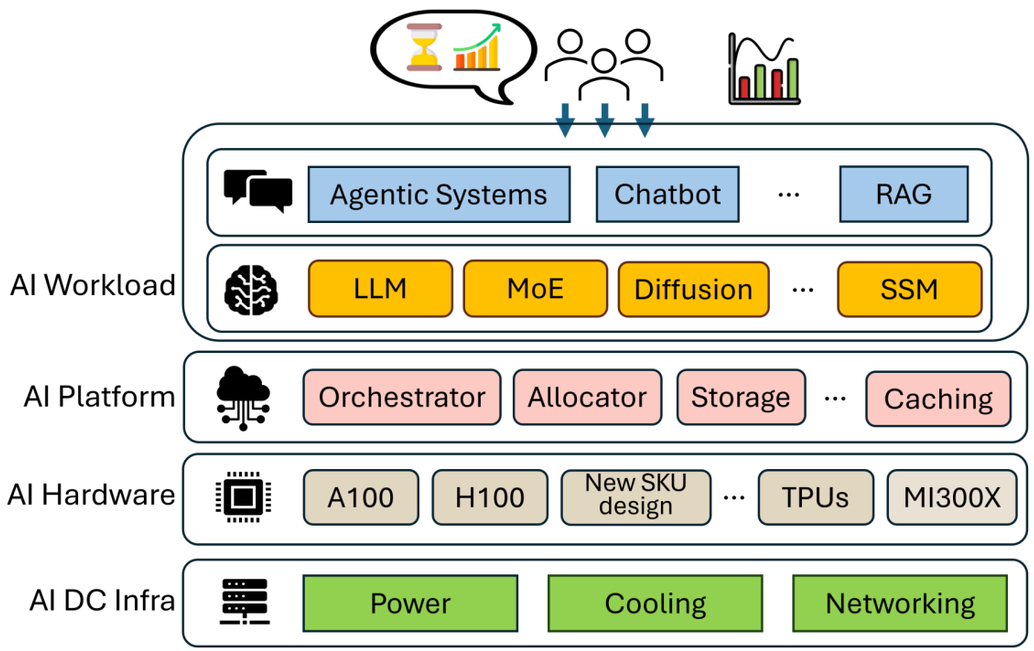
- AI inference is becoming a dominant datacenter workload
- AI workloads and hardware are fast evolving
- Traditional data[illegible]ned for CPUs

Does AI change the economics of datacenters?



Motivation

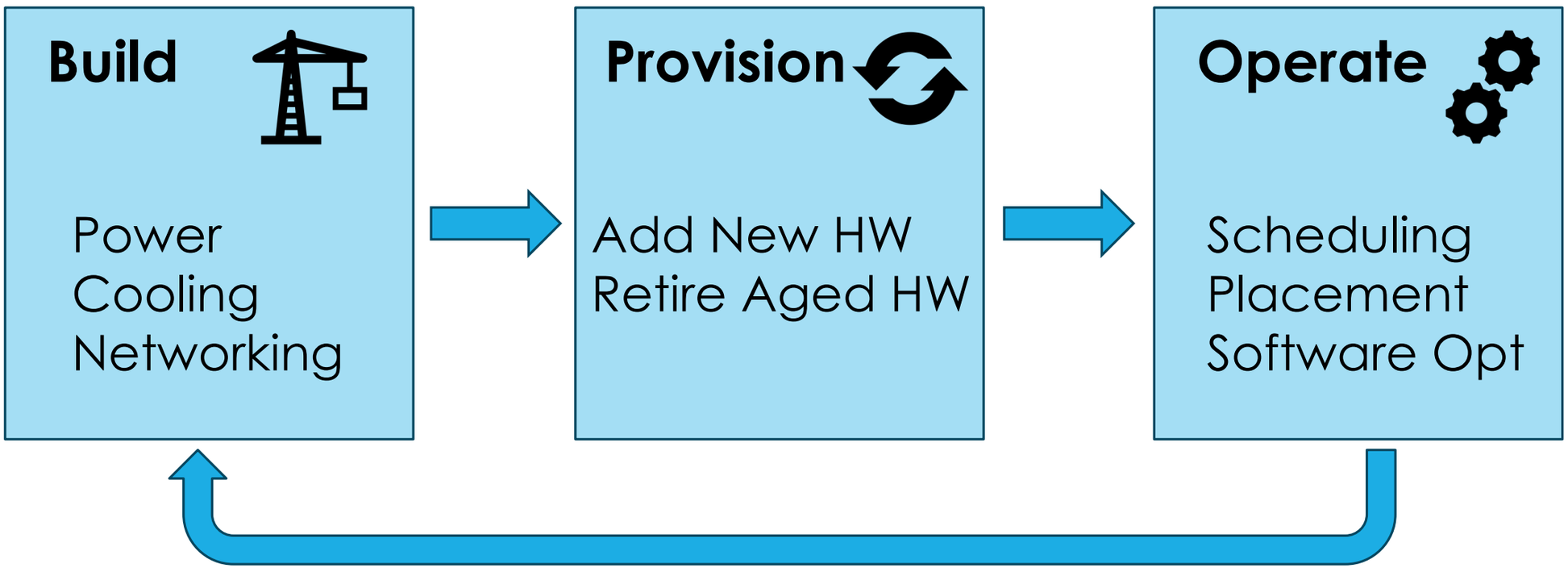
- How should cloud providers build, provision, and operate AI datacenters to minimize long-term TCO?



Contributions

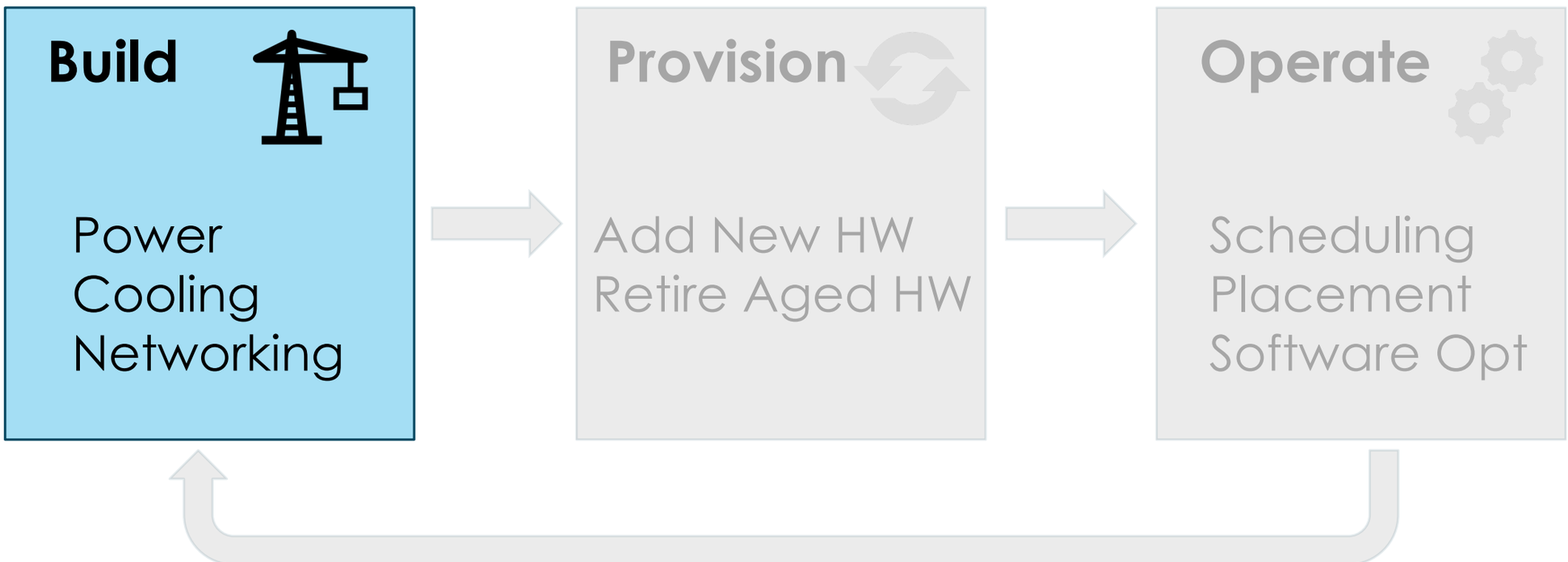
- Rethink the AI datacenter lifecycle across build, provisioning, and operation stages through a TCO-driven framework
- 40% TCO reduction through holistic cross-stage optimization

A Datacenter Lifecycle

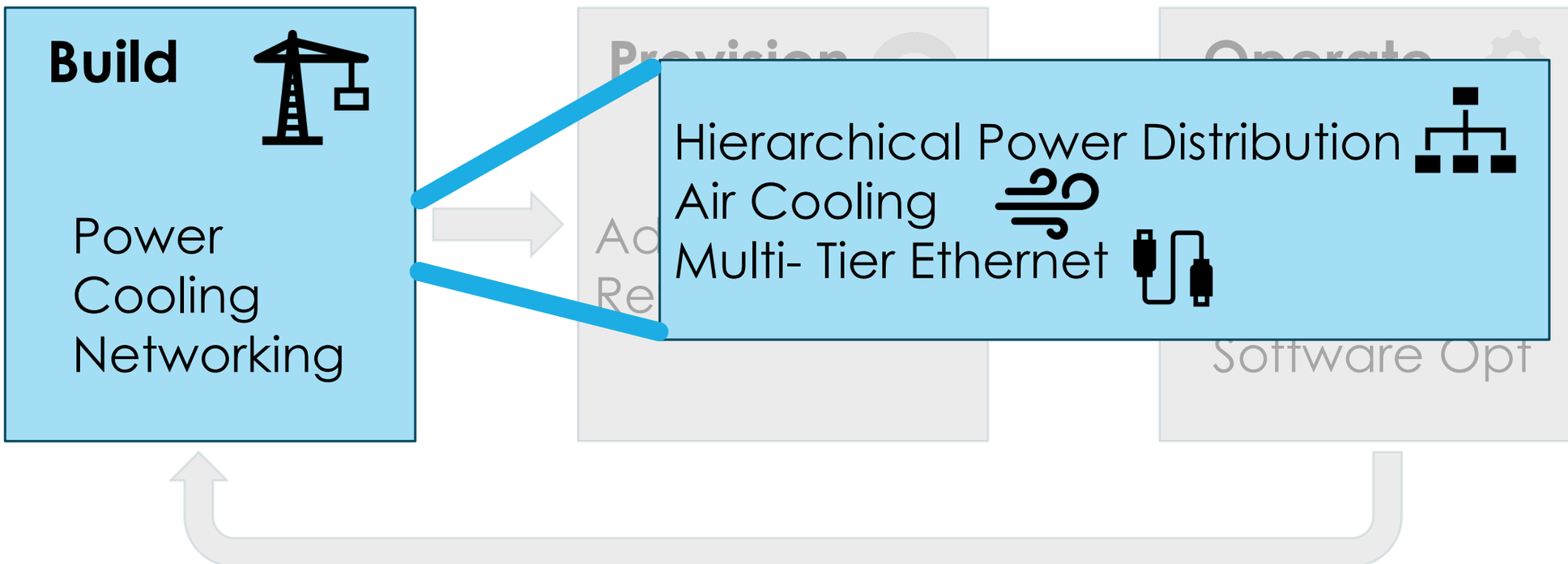


A Datacenter Lifecycle: Traditional Approach

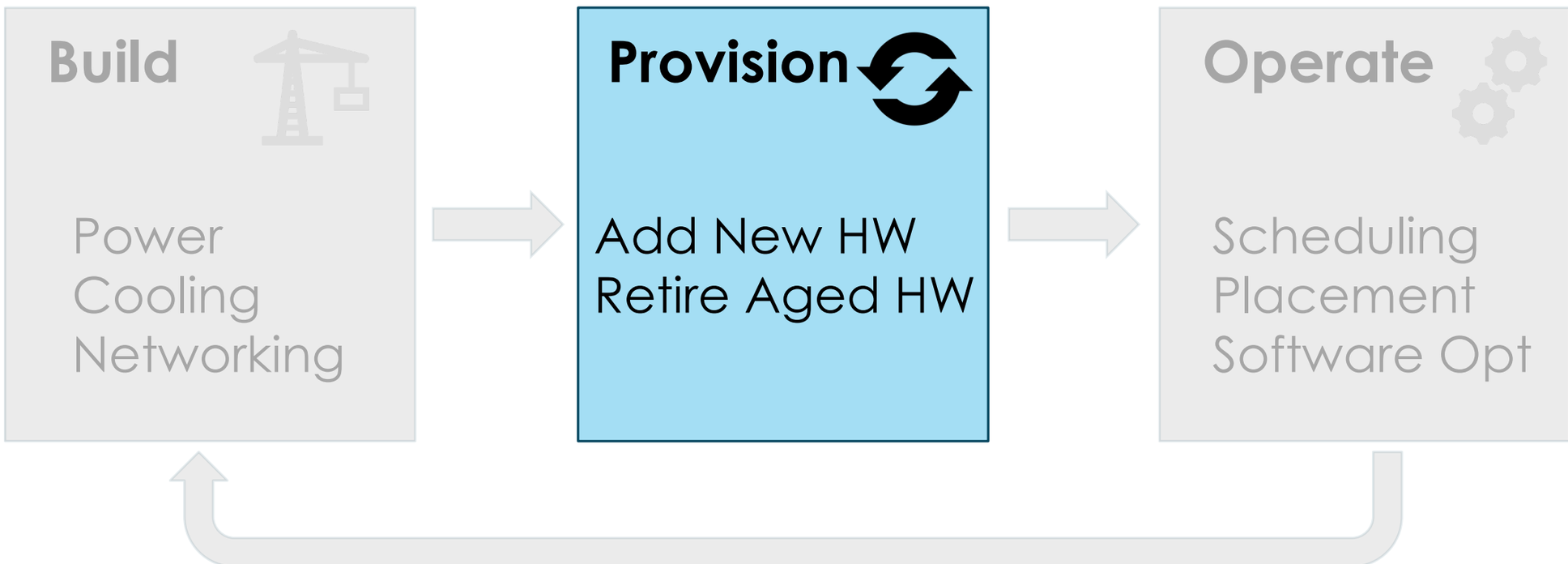
A Datacenter Lifecycle: Traditional Approach



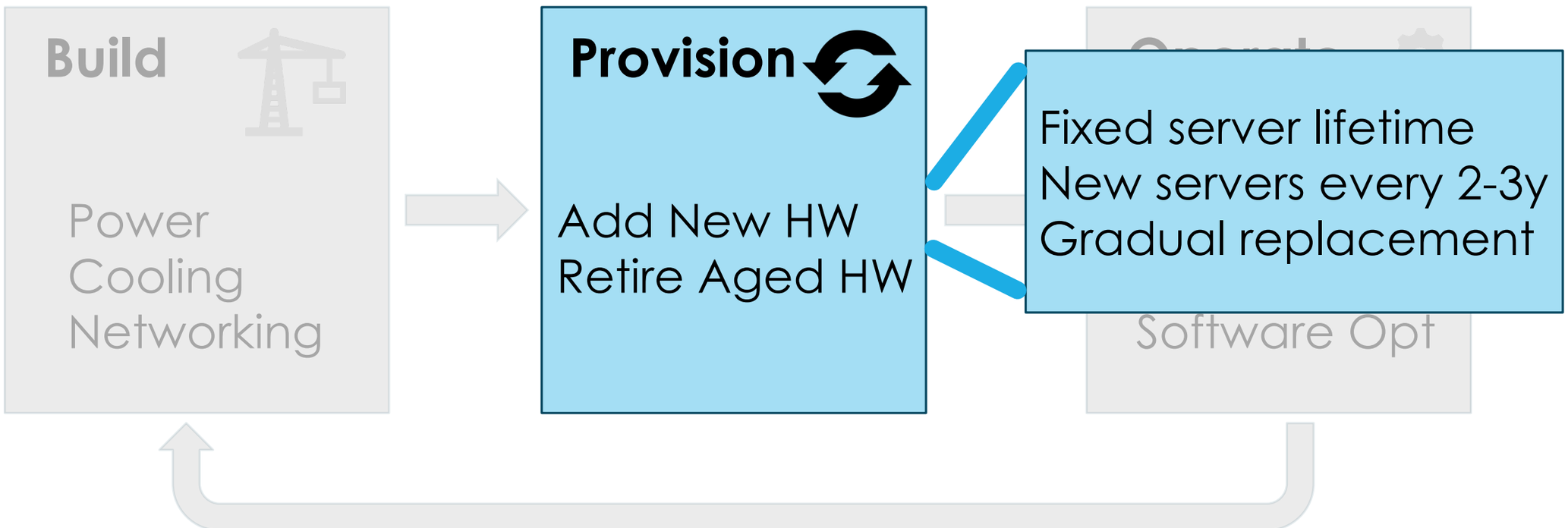
A Datacenter Lifecycle: Traditional Approach



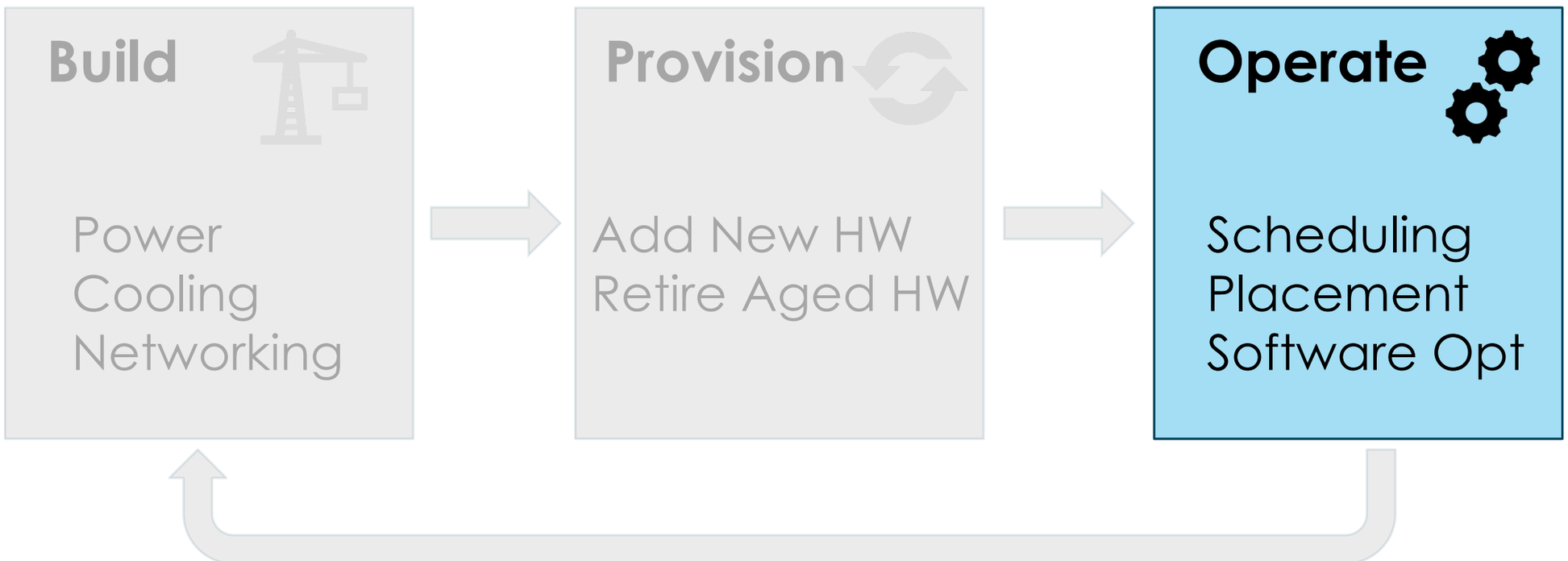
A Datacenter Lifecycle: Traditional Approach



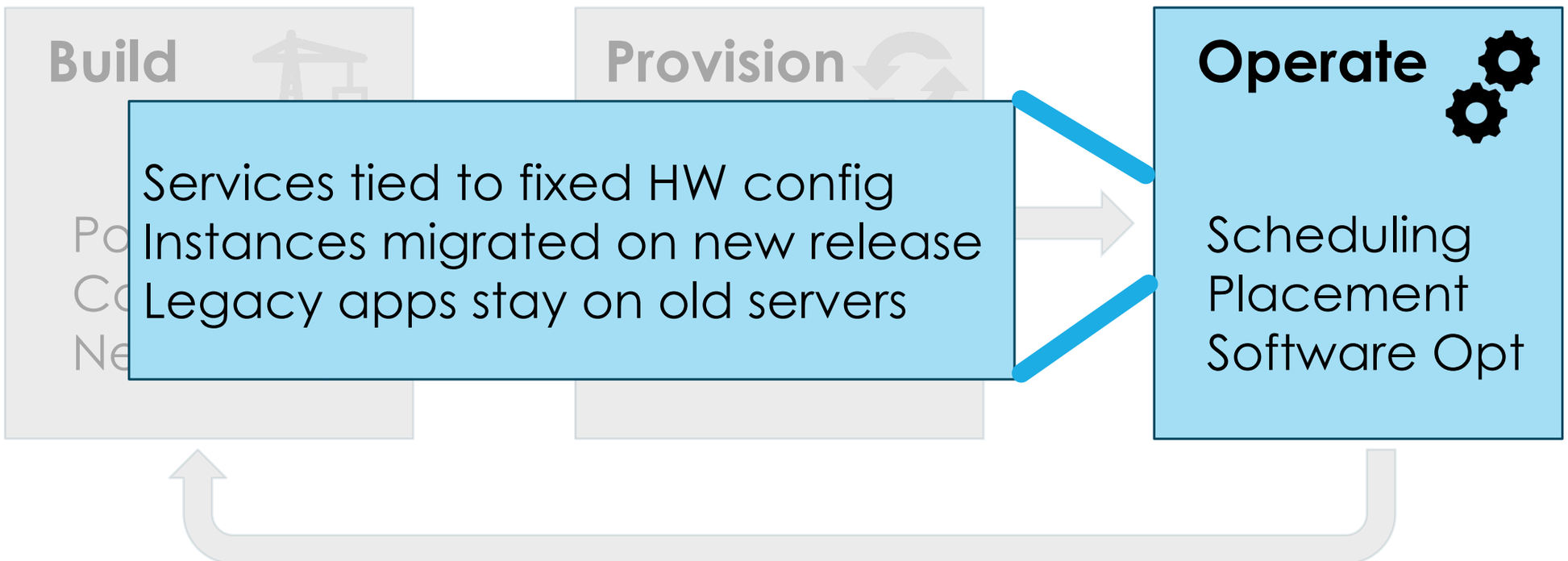
A Datacenter Lifecycle: Traditional Approach



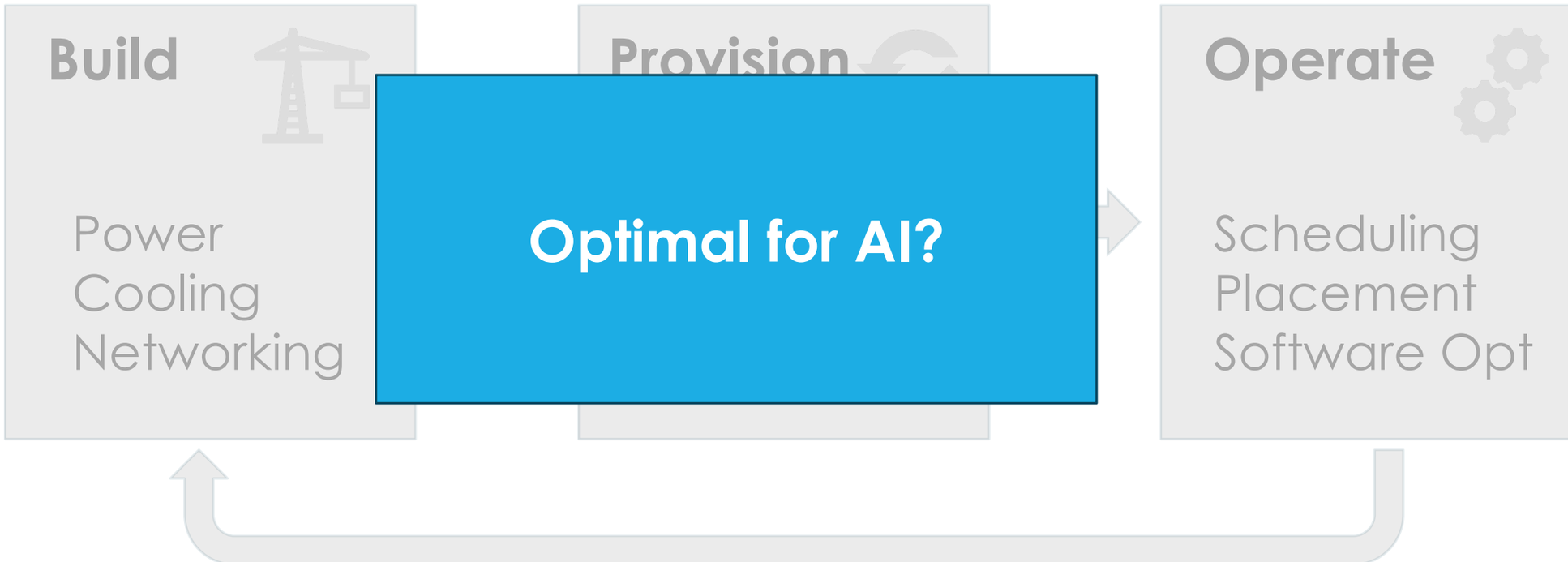
A Datacenter Lifecycle: Traditional Approach



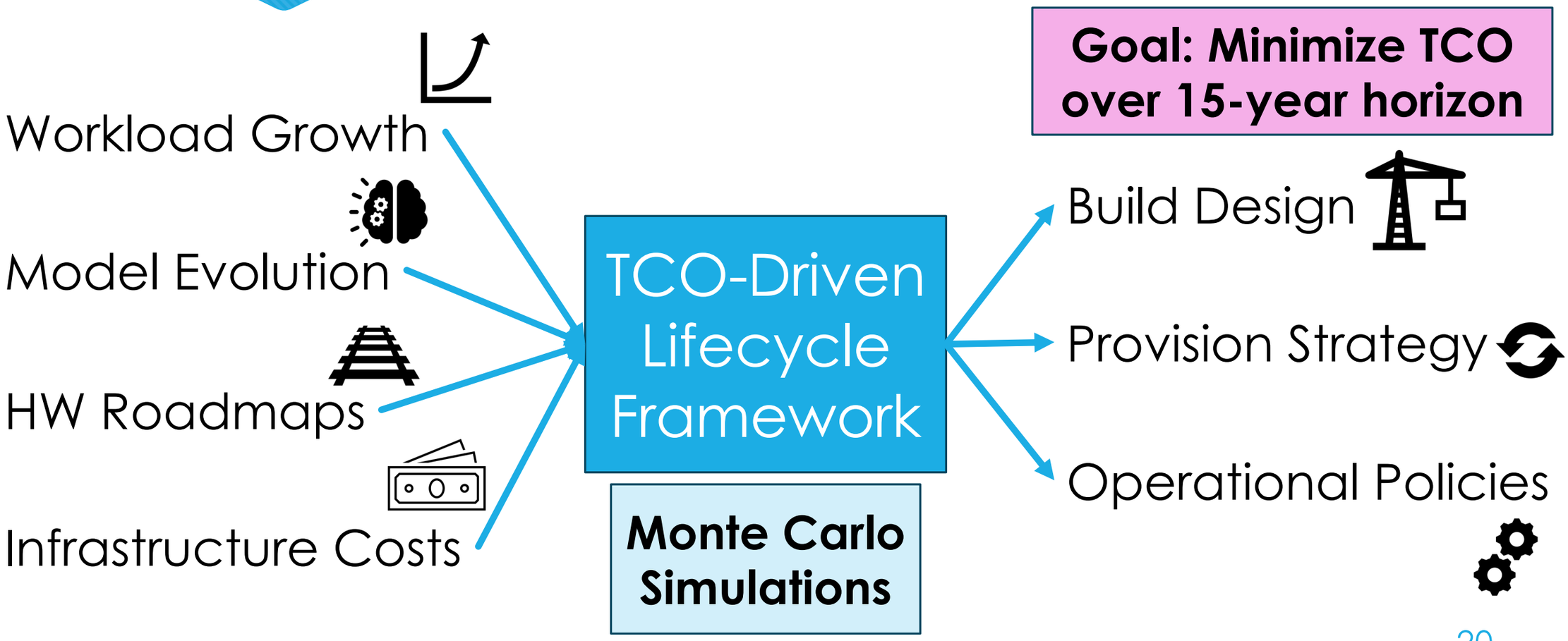
A Datacenter Lifecycle: Traditional Approach



A Datacenter Lifecycle: Traditional Approach



TCO-Driven AI Datacenter Lifecycle Framework

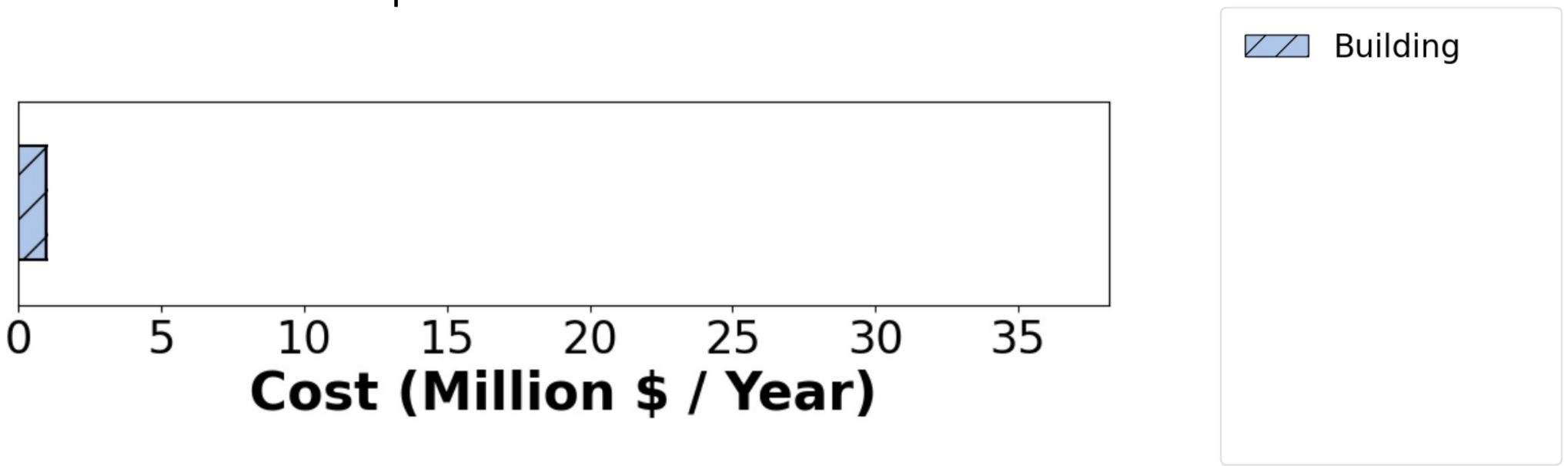


TCO-Driven AI Datacenter Lifecycle Framework

- TCO of an example 10MW datacenter

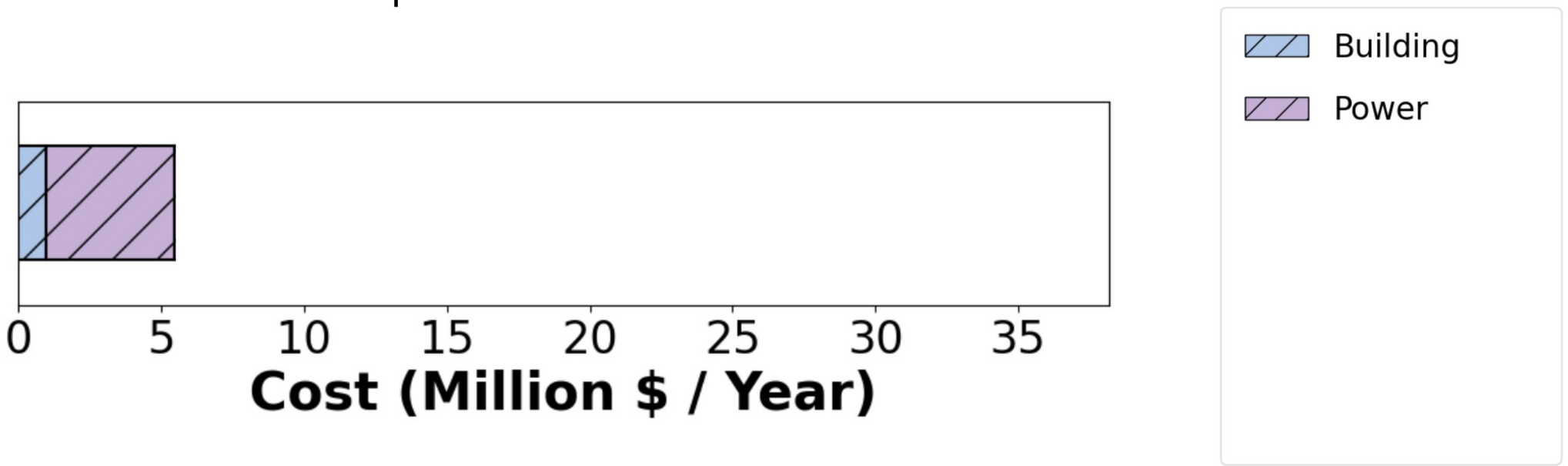
TCO-Driven AI Datacenter Lifecycle Framework

- TCO of an example 10MW datacenter



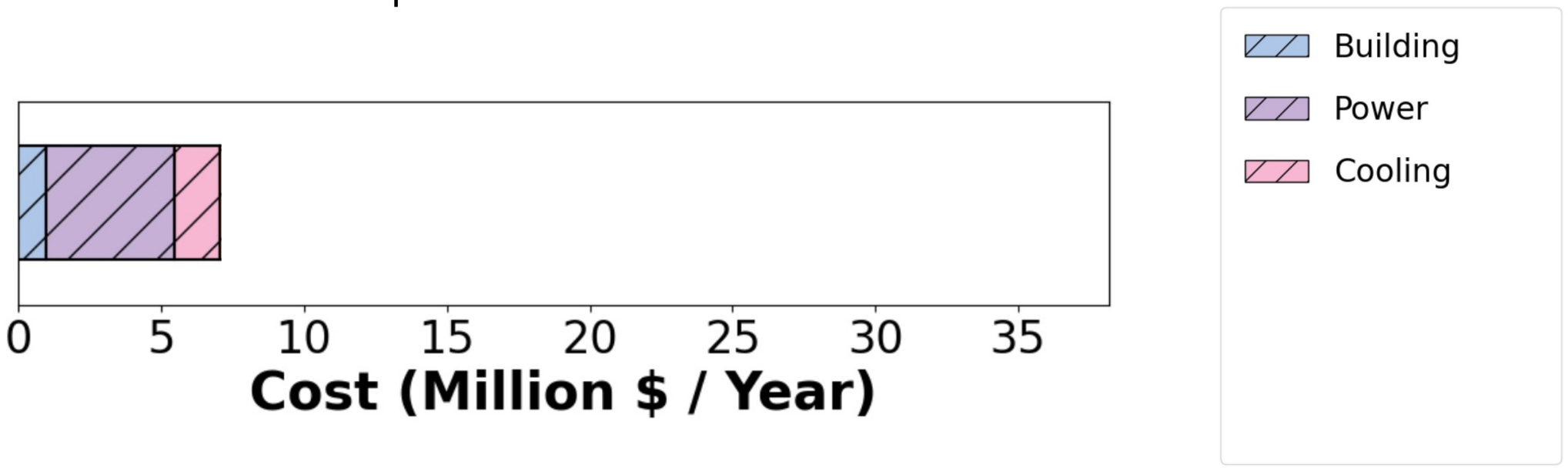
TCO-Driven AI Datacenter Lifecycle Framework

- TCO of an example 10MW datacenter



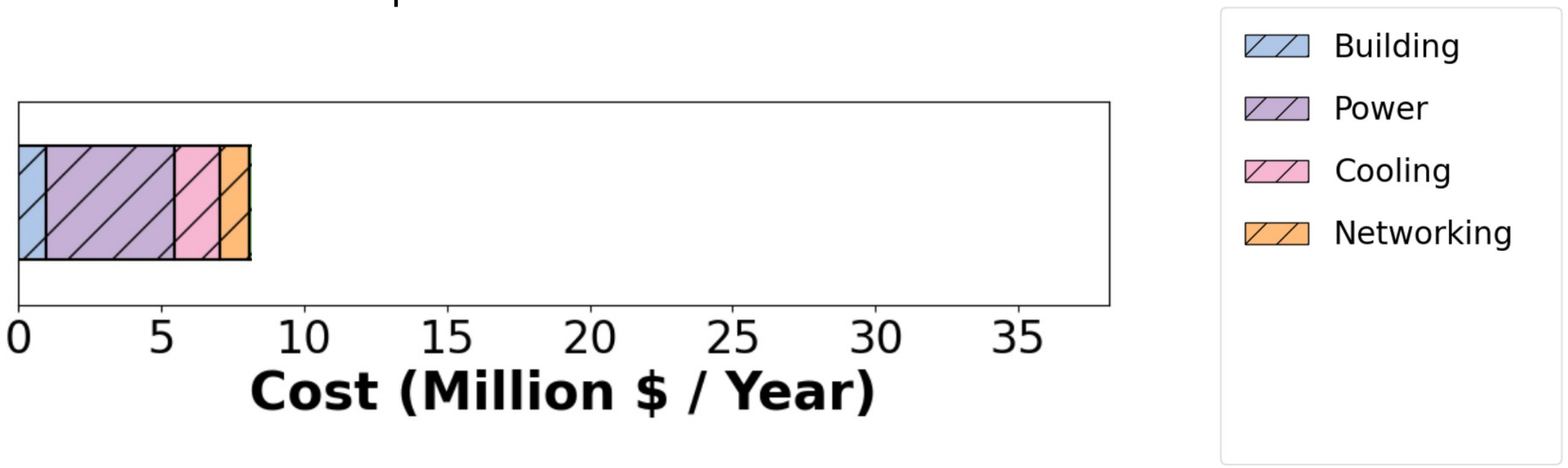
TCO-Driven AI Datacenter Lifecycle Framework

- TCO of an example 10MW datacenter



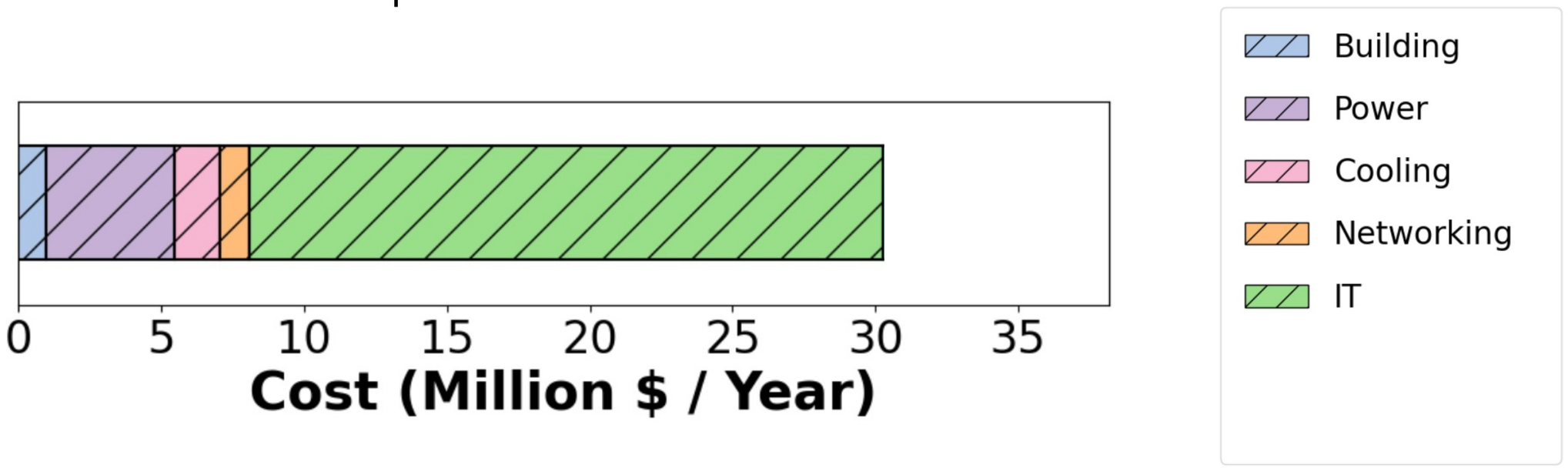
TCO-Driven AI Datacenter Lifecycle Framework

- TCO of an example 10MW datacenter



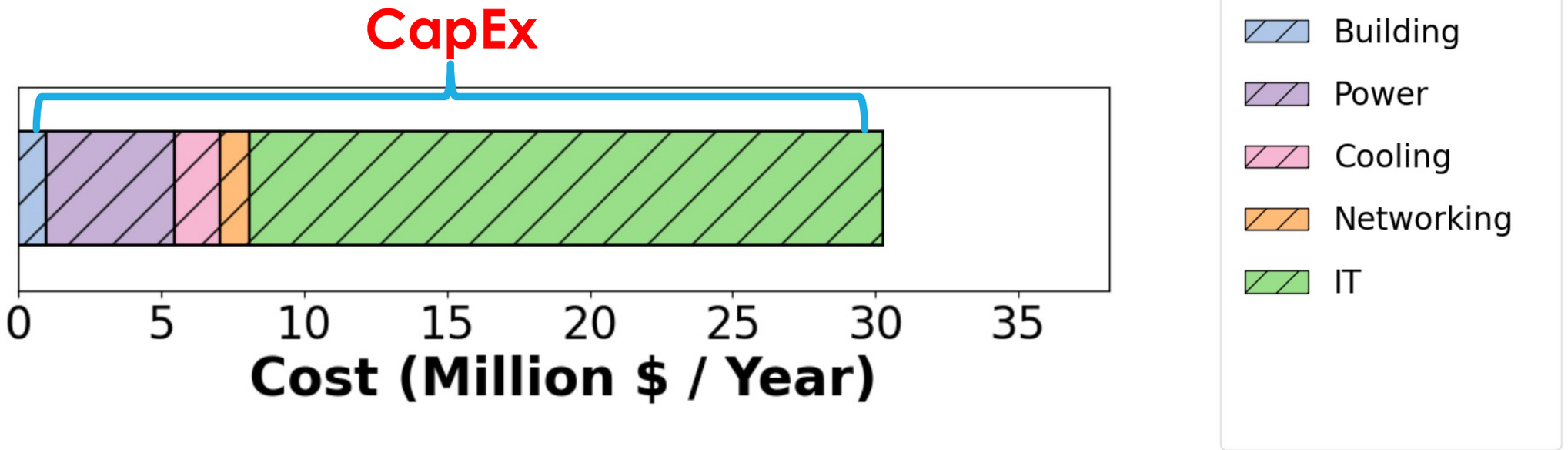
TCO-Driven AI Datacenter Lifecycle Framework

- TCO of an example 10MW datacenter



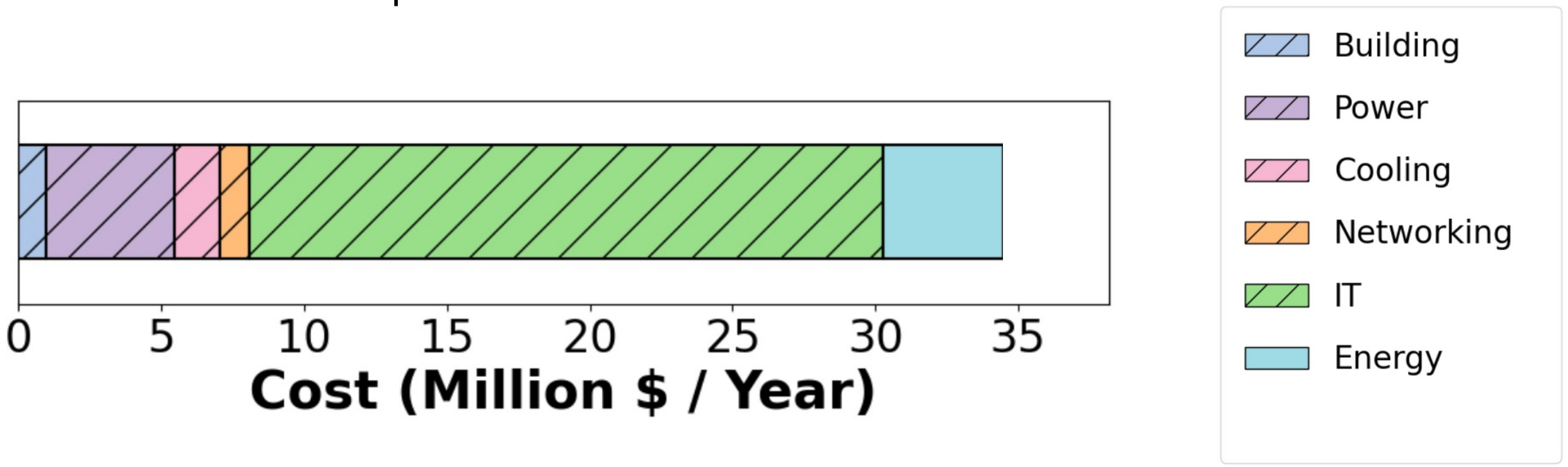
TCO-Driven AI Datacenter Lifecycle Framework

- TCO of an example 10MW datacenter



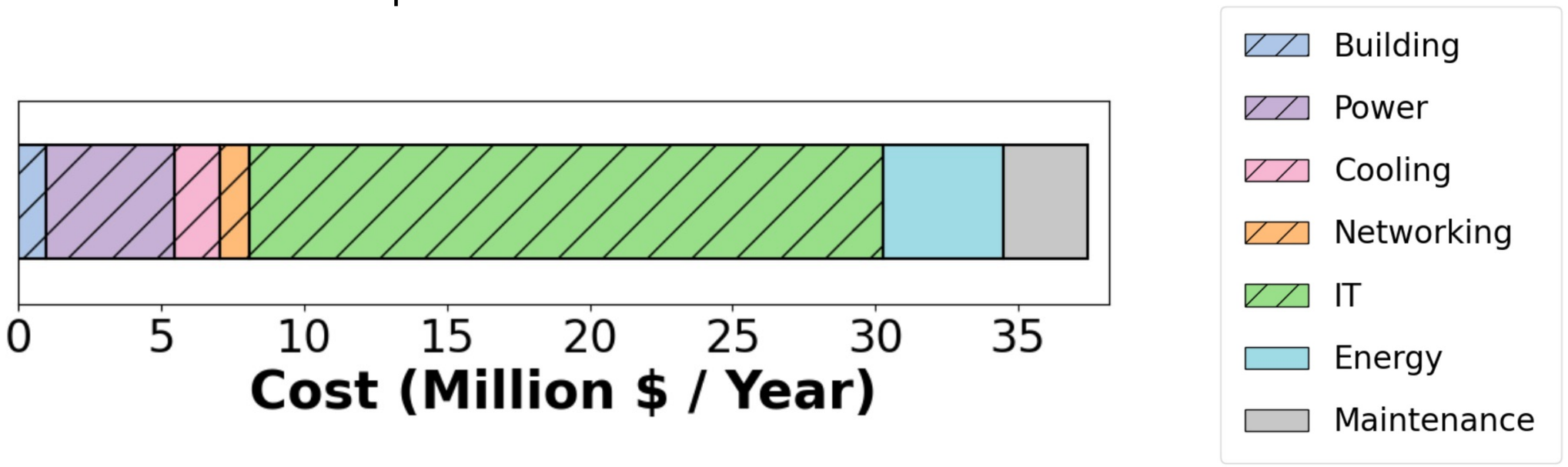
TCO-Driven AI Datacenter Lifecycle Framework

- TCO of an example 10MW datacenter



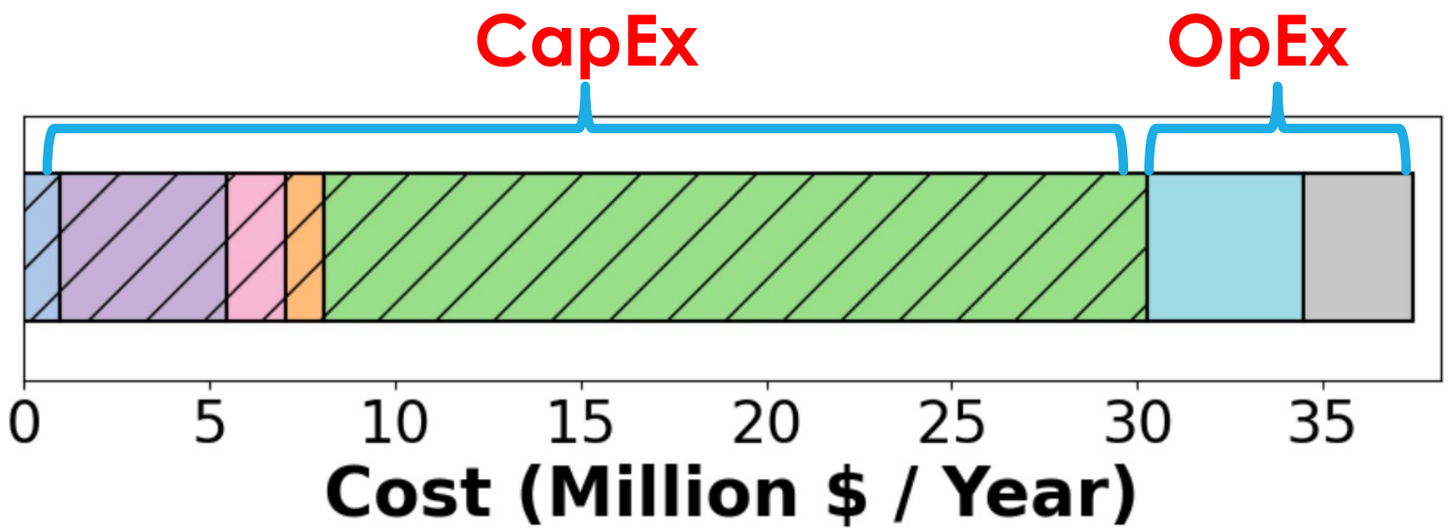
TCO-Driven AI Datacenter Lifecycle Framework

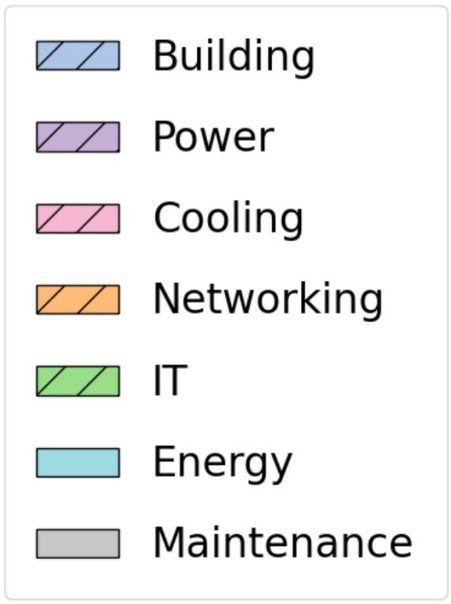
- TCO of an example 10MW datacenter



TCO-Driven AI Datacenter Lifecycle Framework

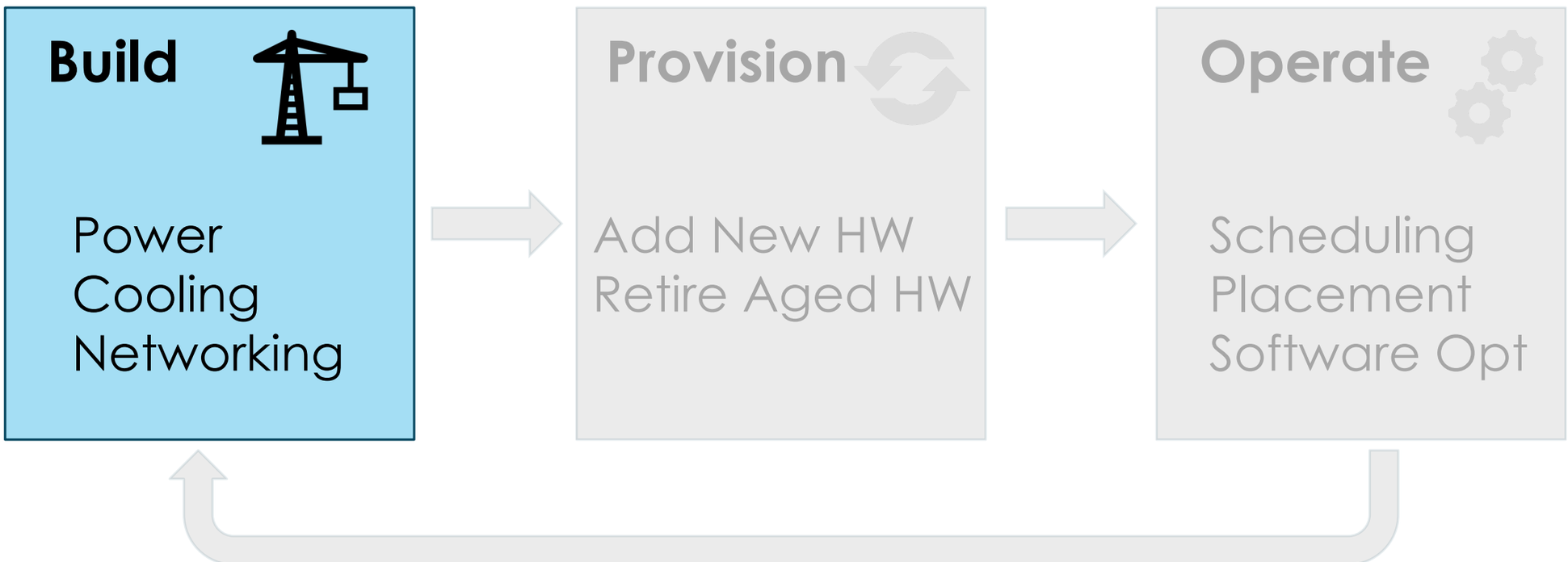
- TCO of an example 10MW datacenter





A Datacenter Lifecycle: Key Findings

A Datacenter Lifecycle: Key Findings

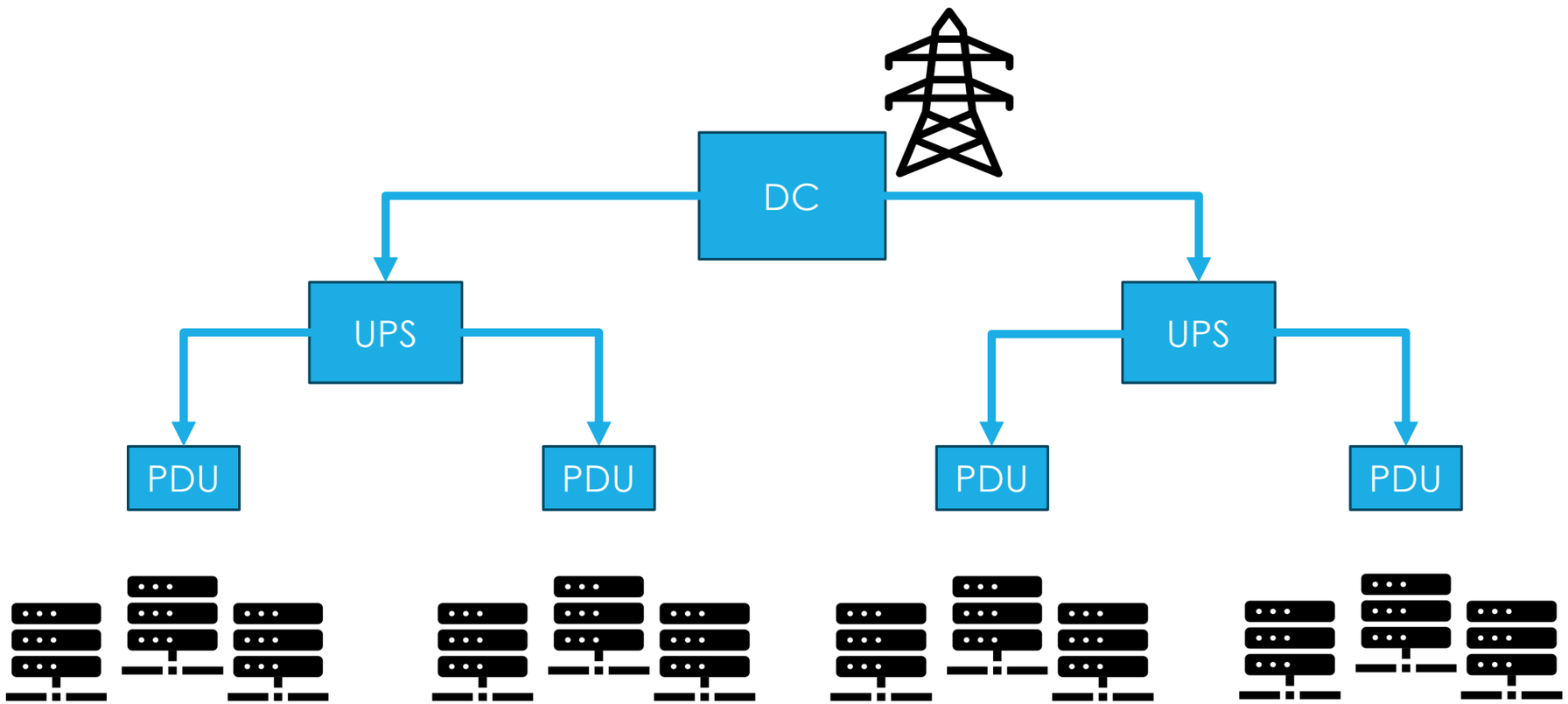


Build-Stage Findings: Hierarchical Power Topology → Fragmentation

- Traditional datacenters: hierarchical power distribution

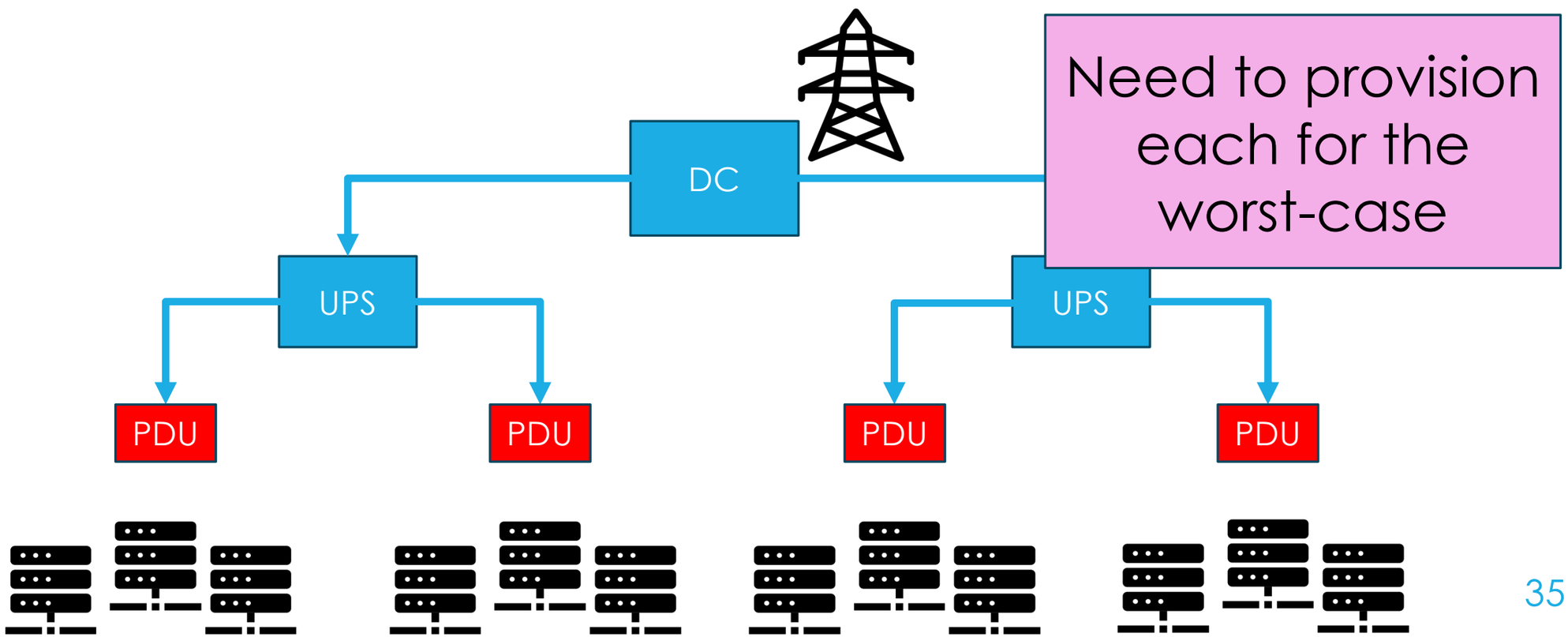
Build-Stage Findings: Hierarchical Power Topology → Fragmentation

- Traditional datacenters: hierarchical power distribution



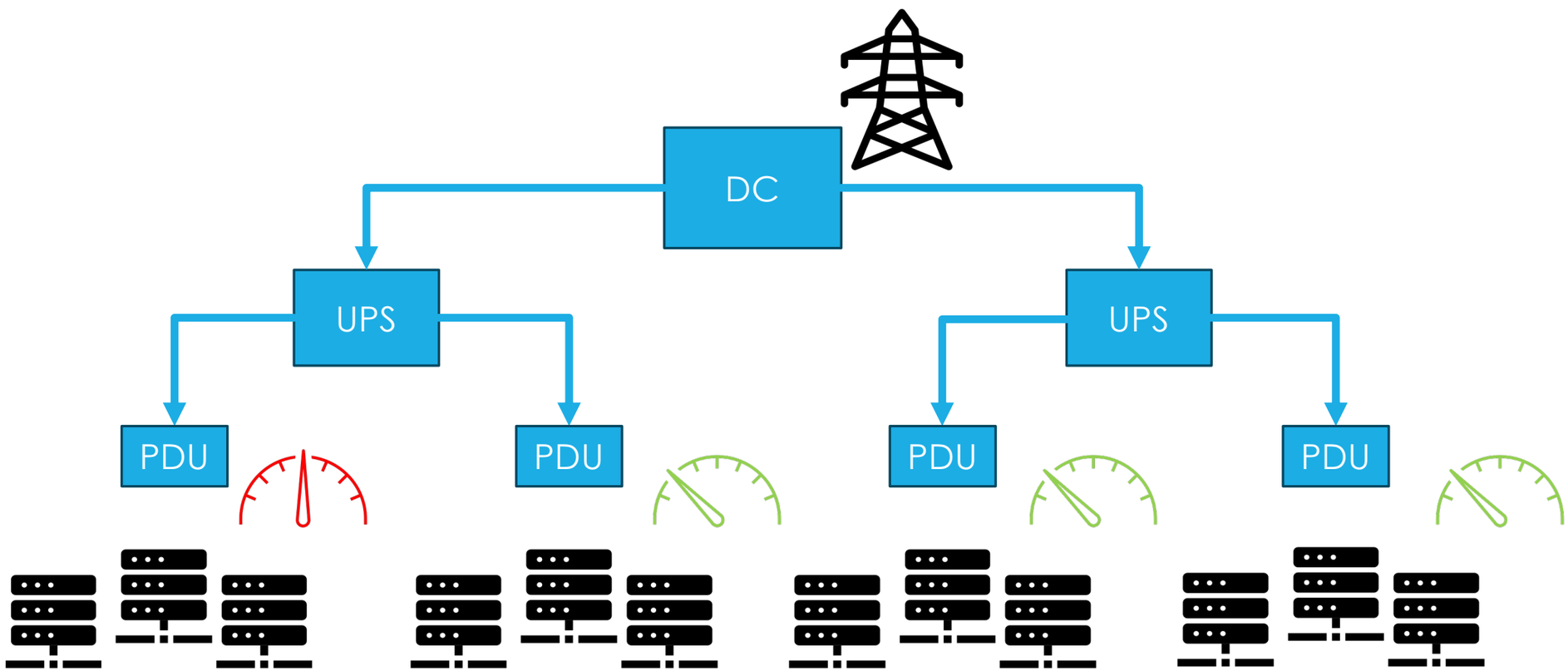
Build-Stage Findings: Hierarchical Power Topology → Fragmentation

- Traditional datacenters: hierarchical power distribution



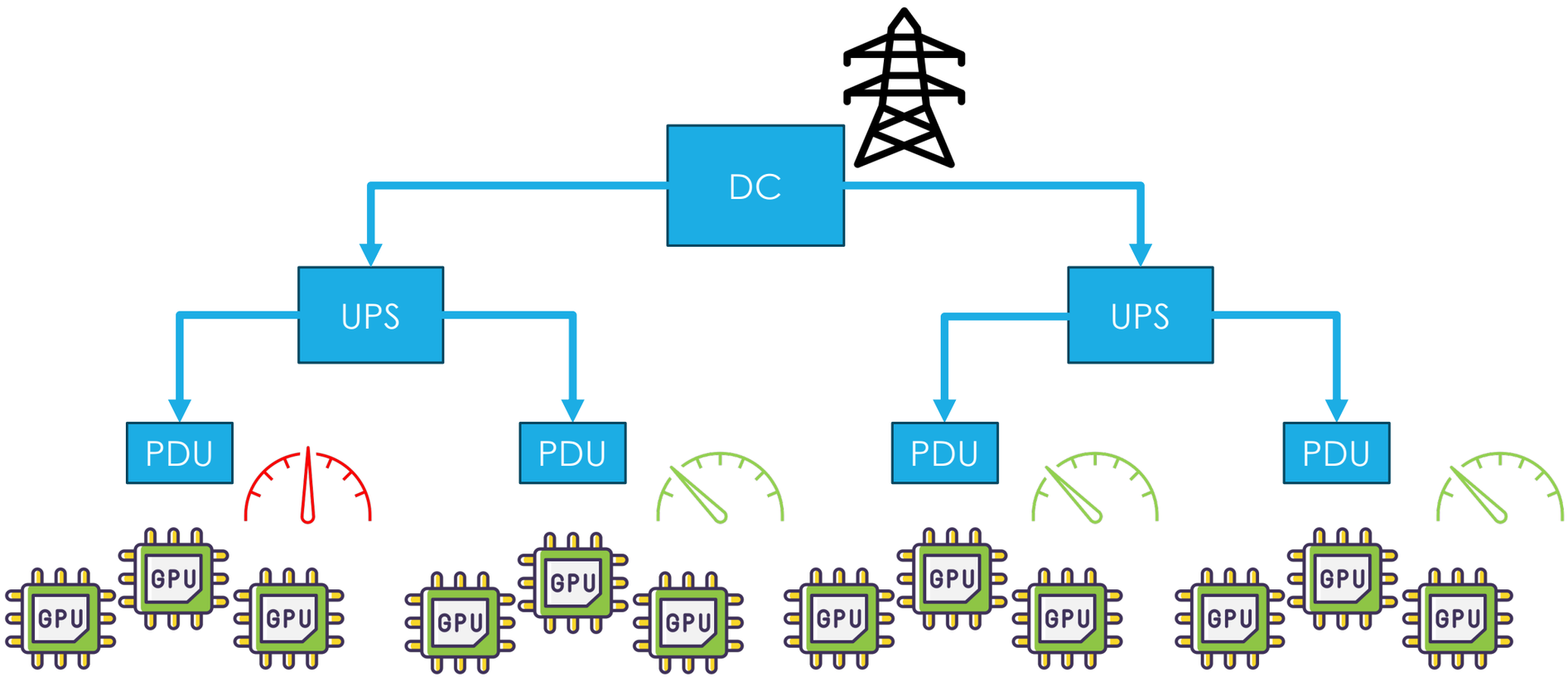
Build-Stage Findings: Hierarchical Power Topology → Fragmentation

- Traditional datacenters: hierarchical power distribution



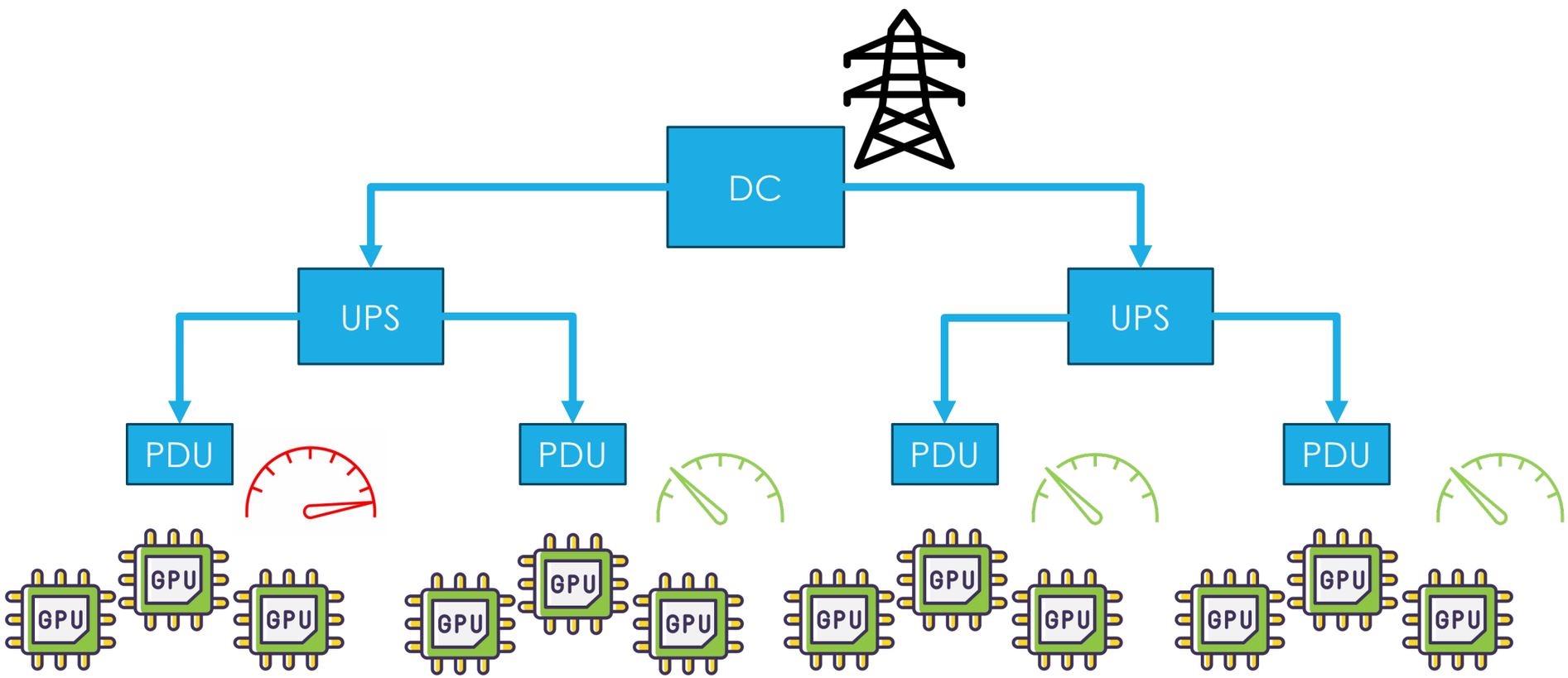
Build-Stage Findings: Hierarchical Power Topology → Fragmentation

- Traditional datacenters: hierarchical power distribution



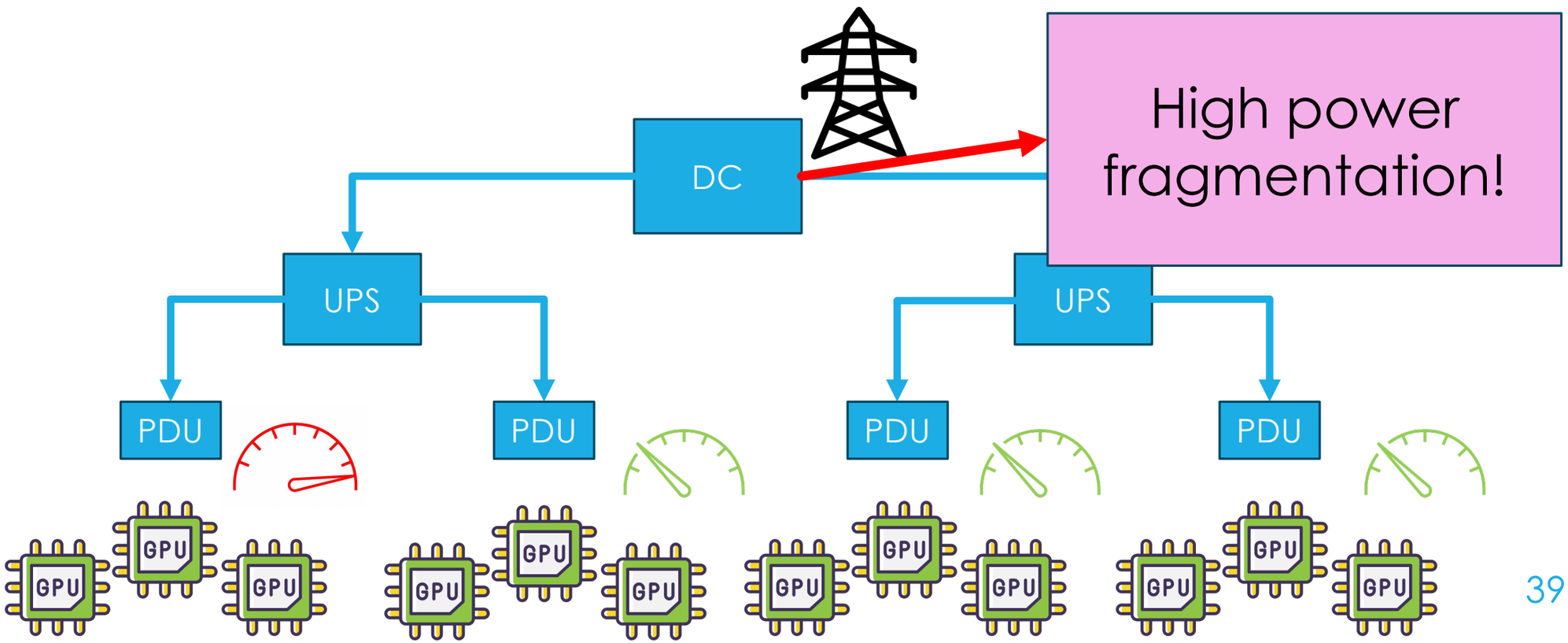
Build-Stage Findings: Hierarchical Power Topology → Fragmentation

- Traditional datacenters: hierarchical power distribution



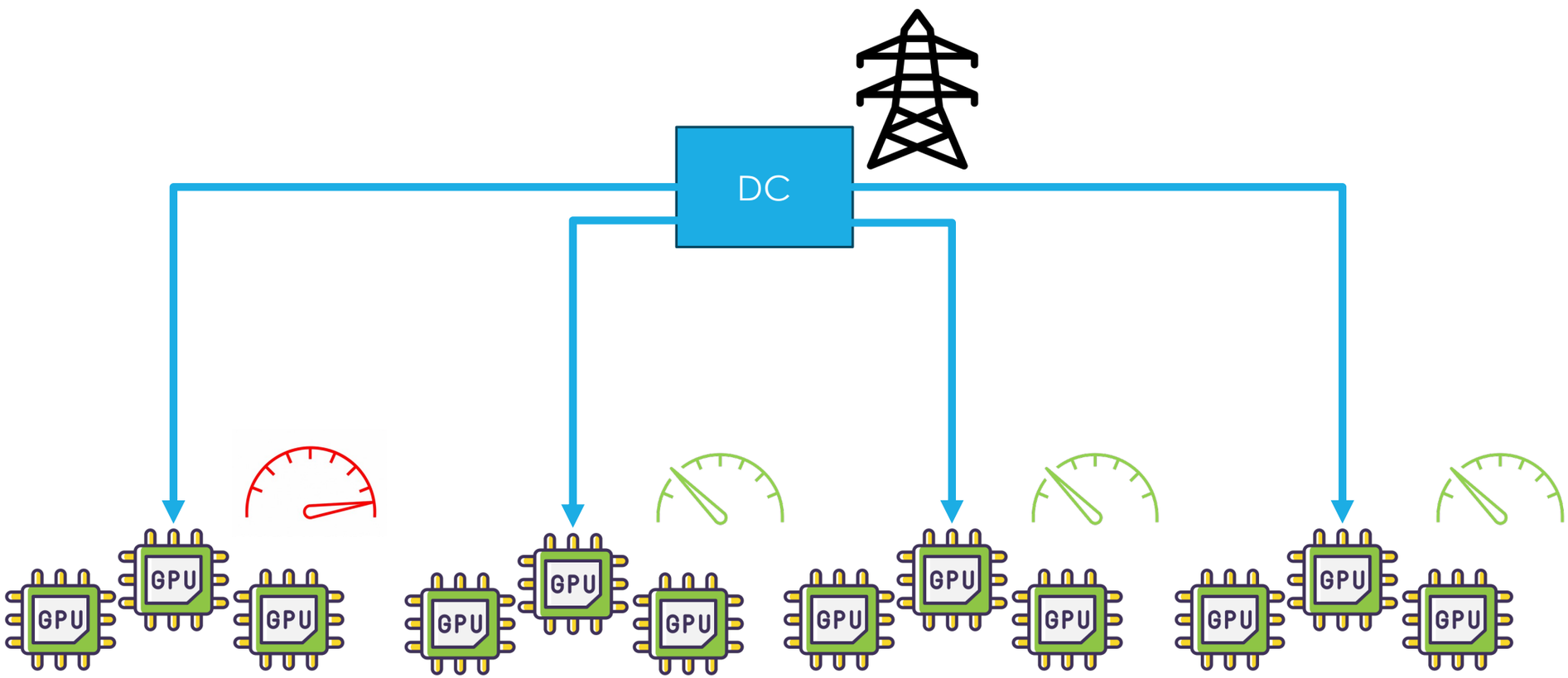
Build-Stage Findings: Hierarchical Power Topology → Fragmentation

- Traditional datacenters: hierarchical power distribution



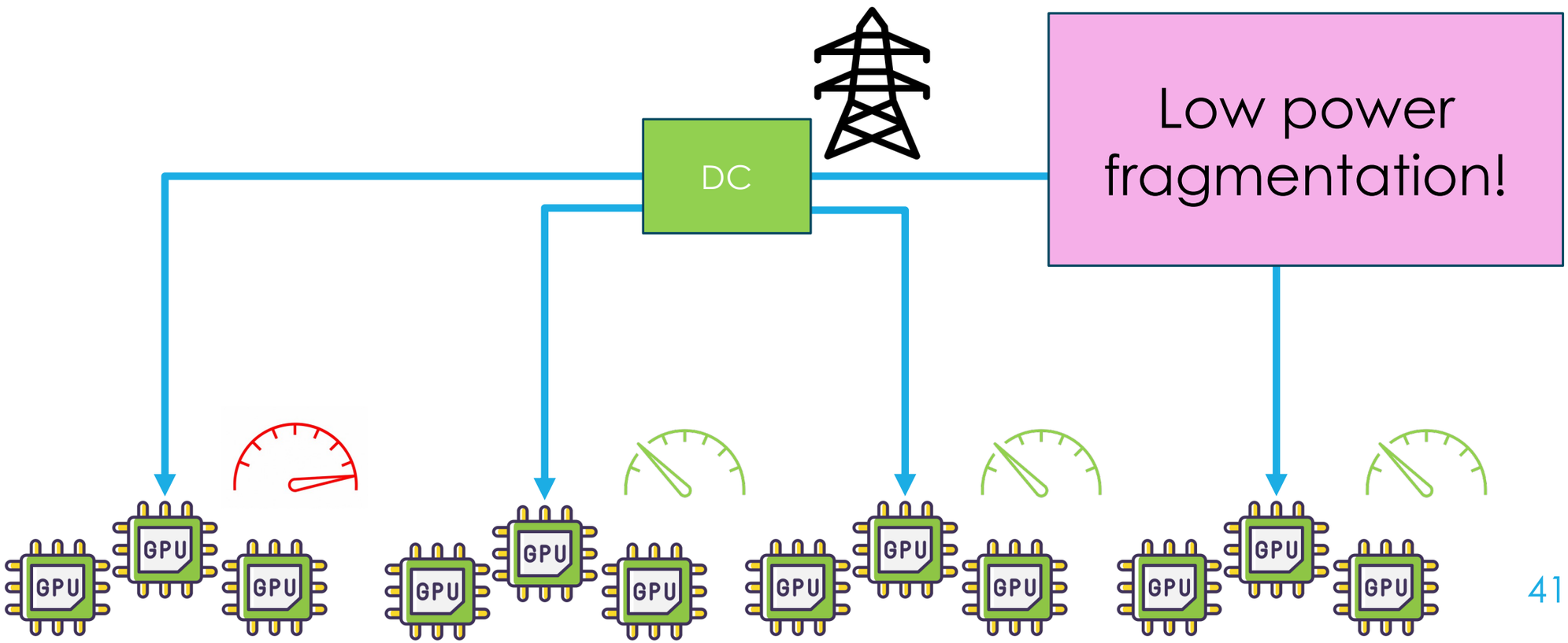
Build-Stage Findings: Need for a Flat Power Topology

- AI datacenters: need flatter power distribution topologies



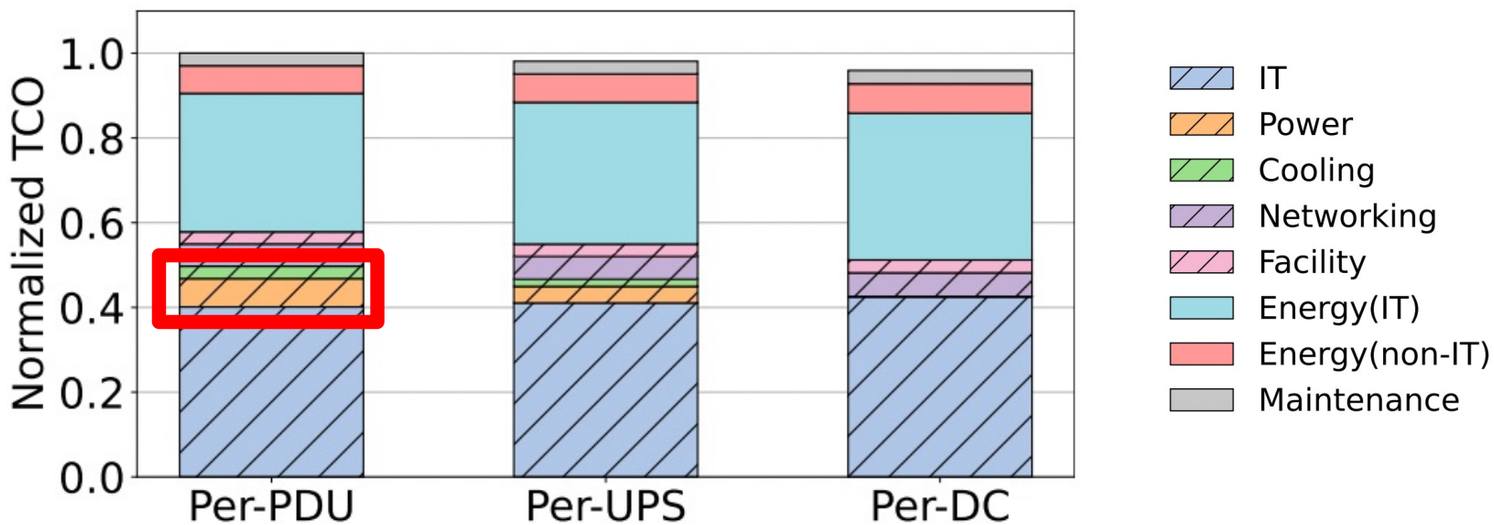
Build-Stage Findings: Need for a Flat Power Topology

- AI datacenters: need flatter power distribution topologies

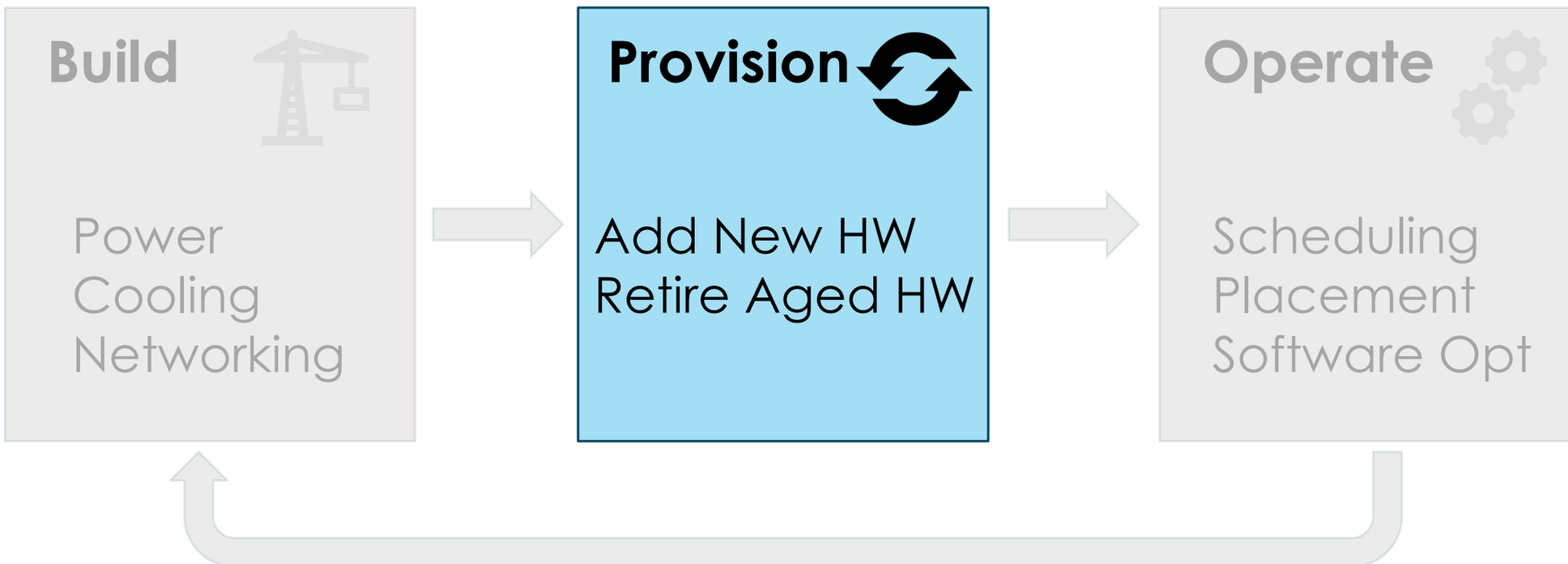


Build-Stage Findings: Need for a Flat Power Topology

- AI datacenters: need flatter power distribution topologies



A Datacenter Lifecycle: Key Findings



Provision-Stage Findings: Traditional Policies Work Well for CPU Datacenters

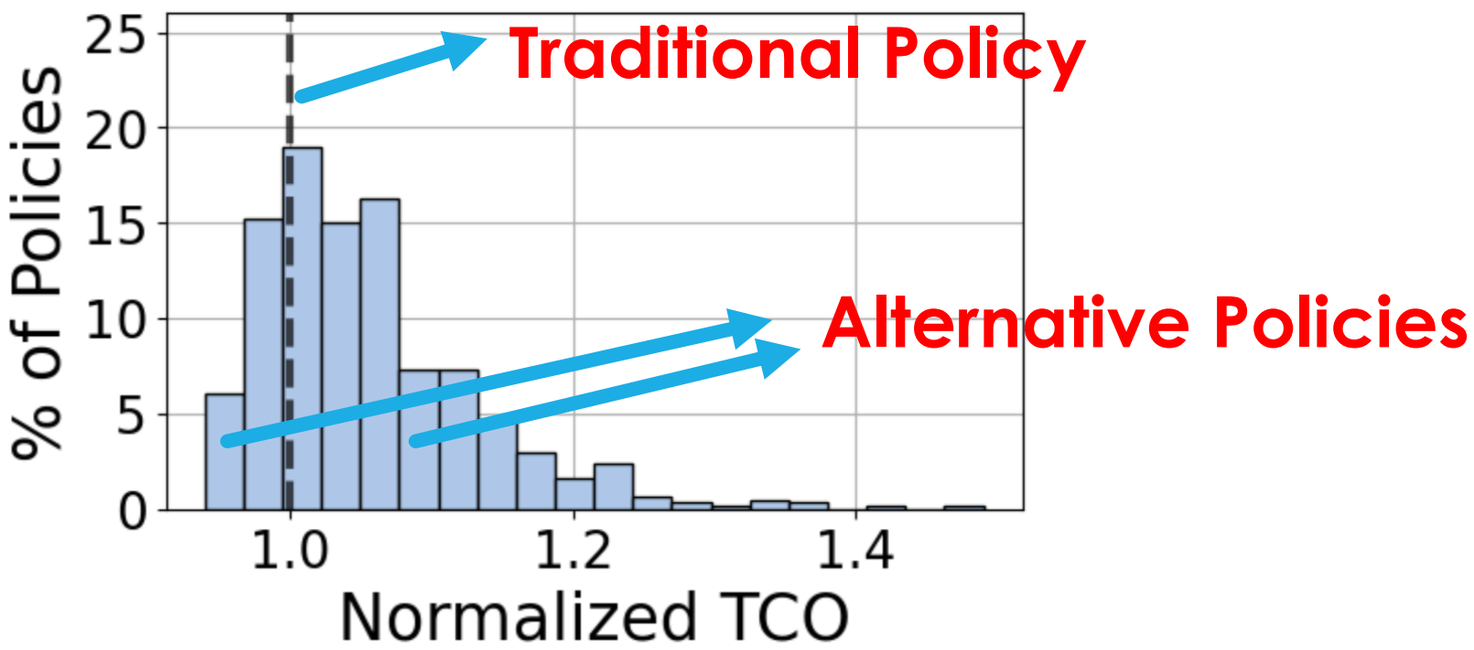
- Traditional policy: steady CPU refresh cycle with 5y of server

Provision-Stage Findings: Traditional Policies Work Well for CPU Datacenters

- Traditional policy: steady CPU refresh cycle with 5y of server
- Alternative policies: extend/shorten lifetime, skip a generation

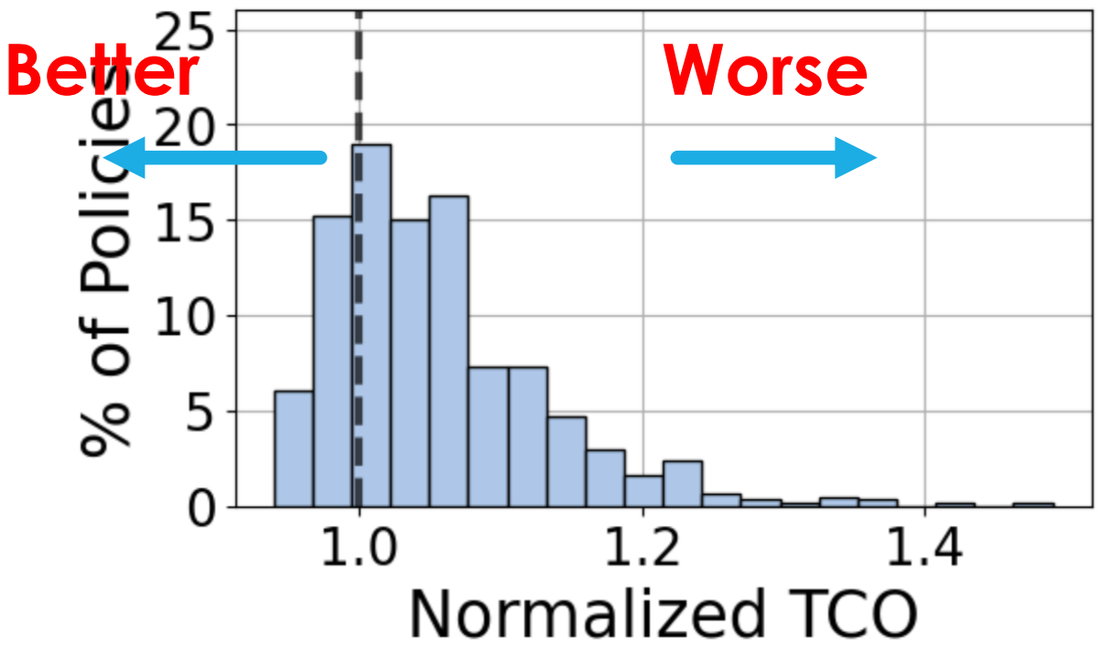
Provision-Stage Findings: Traditional Policies Work Well for CPU Datacenters

- Traditional policy: steady CPU refresh cycle with 5y of server
- Alternative policies: extend/shorten lifetime, skip a generation



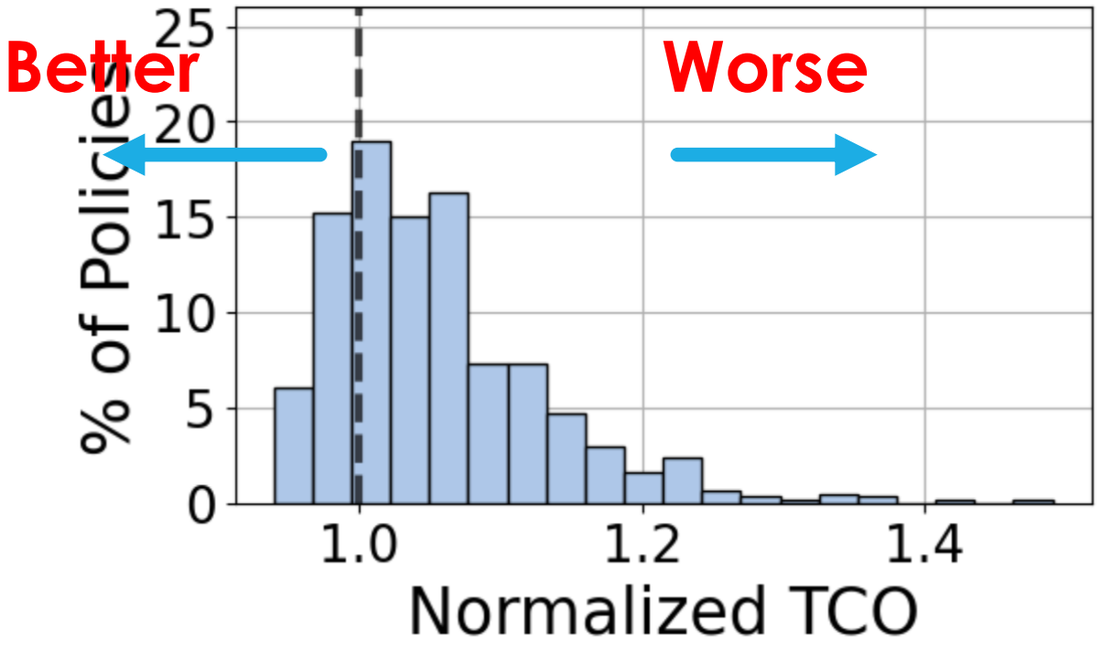
Provision-Stage Findings: Traditional Policies Work Well for CPU Datacenters

- Traditional policy: steady CPU refresh cycle with 5y of server
- Alternative policies: extend/shorten lifetime, skip a generation



Provision-Stage Findings: Traditional Policies Work Well for CPU Datacenters

- Traditional policy: steady CPU refresh cycle with 5y of server
- Alternative policies: extend/shorten lifetime, skip a generation



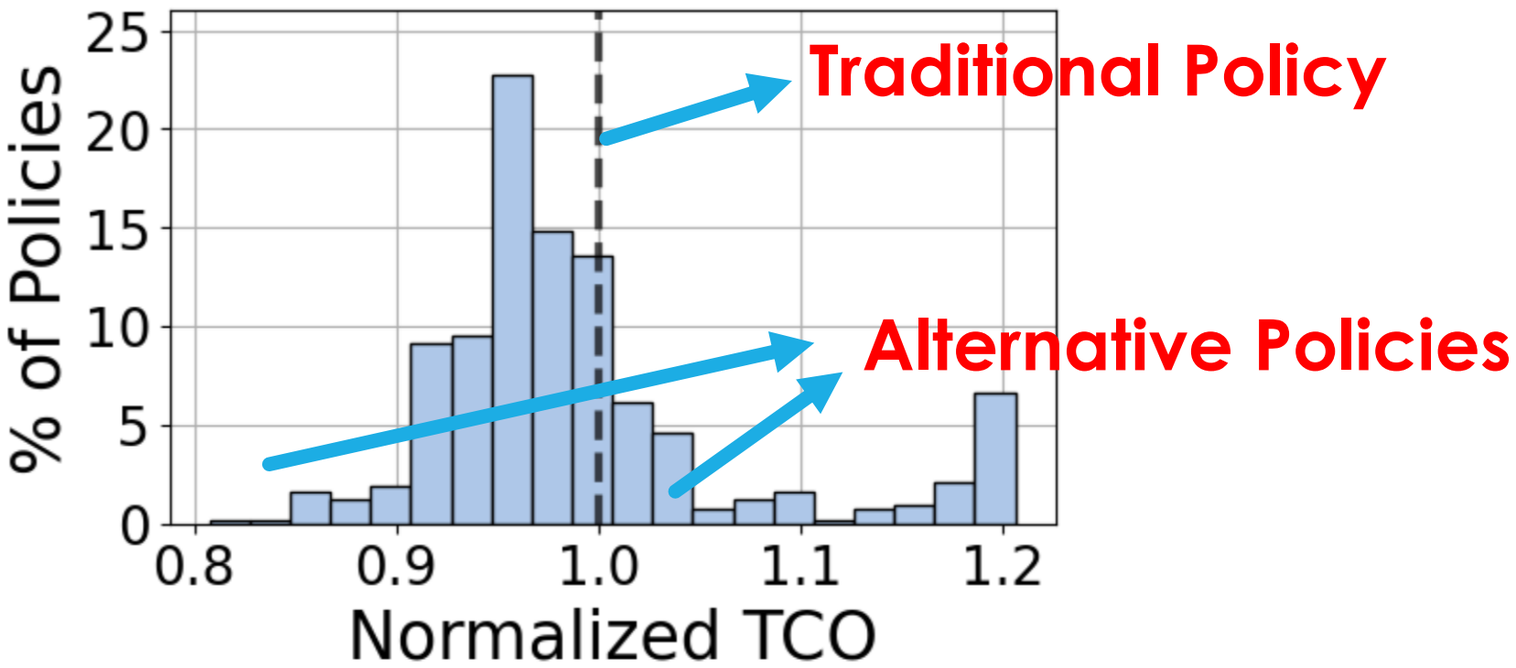
Baseline works well: most alternatives have higher TCO

Provision-Stage Findings: Traditional Policies Fail for AI Datacenters

- The dynamics of AI accelerators and workloads differ substantially from those of general-purpose datacenters

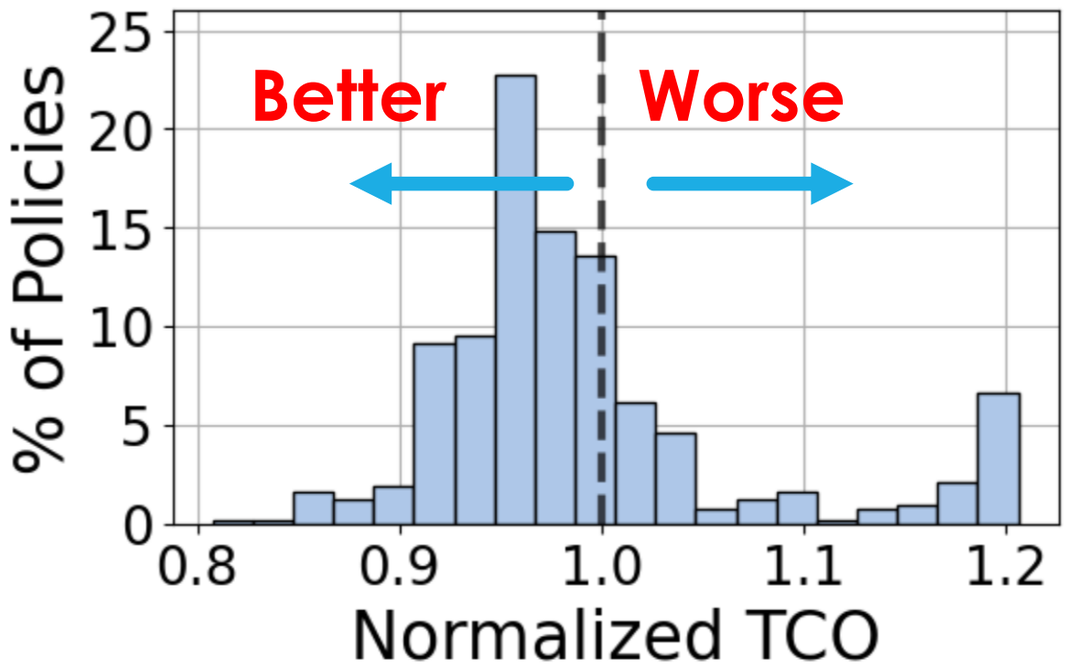
Provision-Stage Findings: Traditional Policies Fail for AI Datacenters

- The dynamics of AI accelerators and workloads differ substantially from those of general-purpose datacenters



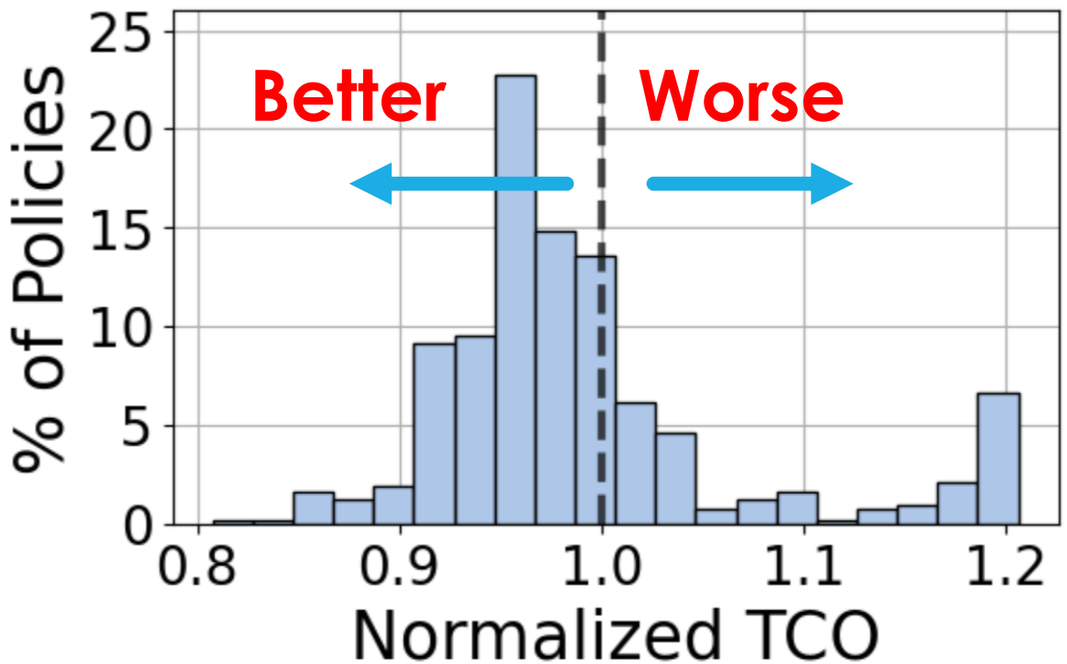
Provision-Stage Findings: Traditional Policies Fail for AI Datacenters

- The dynamics of AI accelerators and workloads differ substantially from those of general-purpose datacenters



Provision-Stage Findings: Traditional Policies Fail for AI Datacenters

- The dynamics of AI accelerators and workloads differ substantially from those of general-purpose datacenters



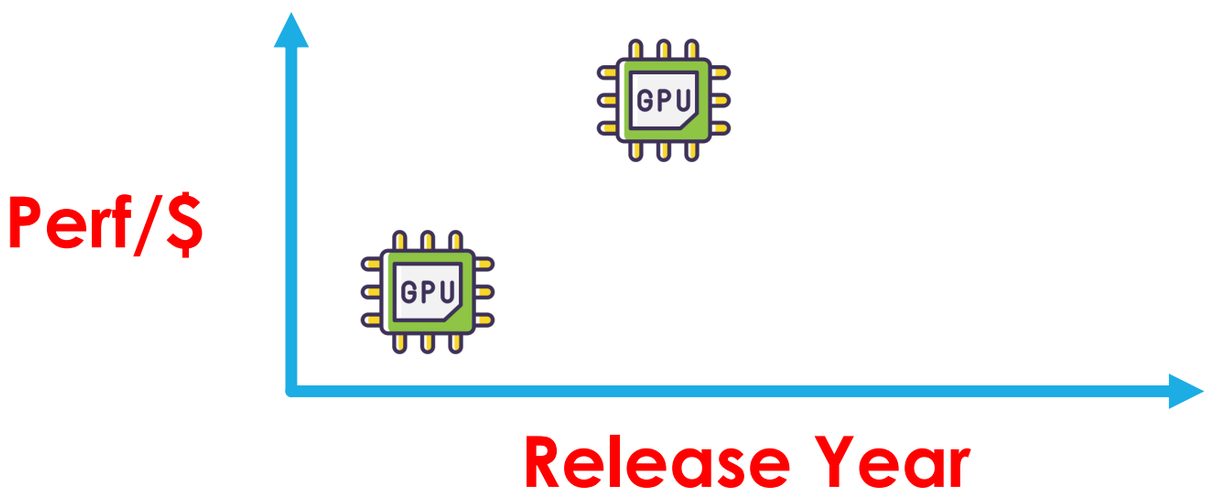
Baseline doesn't work: many alternatives have lower TCO

Provision-Stage Findings: AI Hardware Affects Provisioning Policies

- Hardware-driven refresh signals:

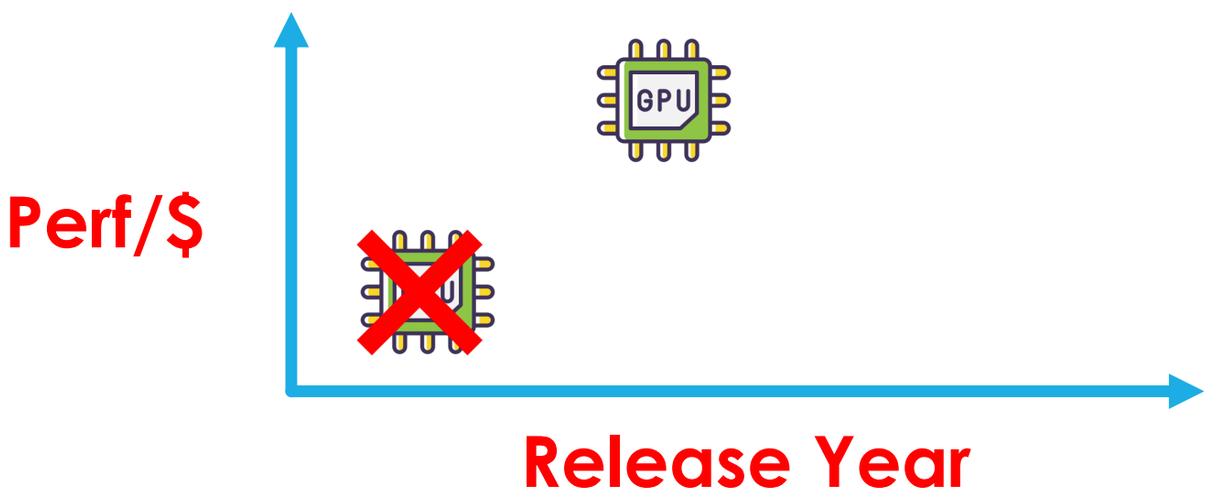
Provision-Stage Findings: AI Hardware Affects Provisioning Policies

- Hardware-driven refresh signals:
  - Newer is much better (H100 vs A100)



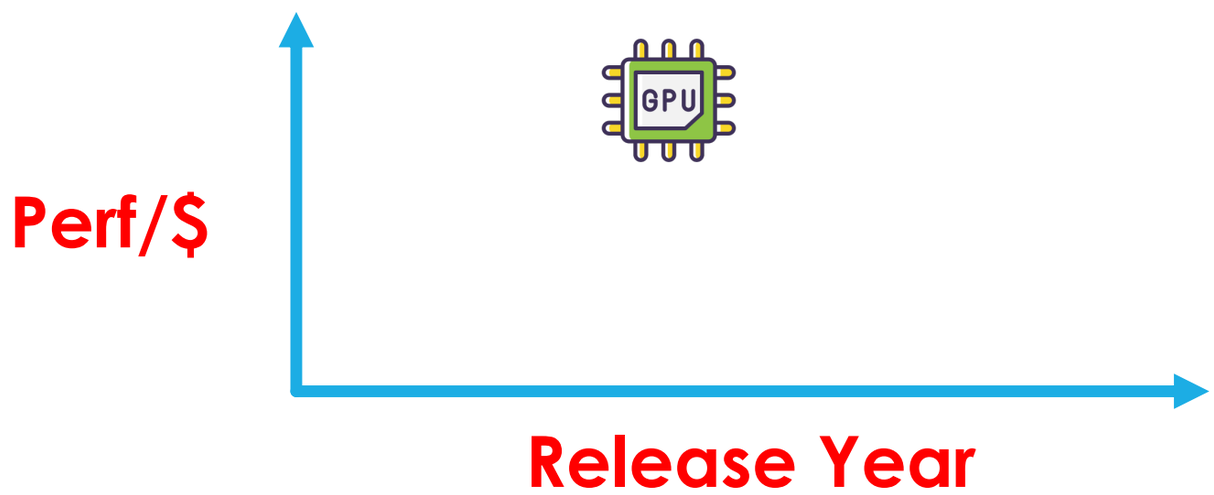
Provision-Stage Findings: AI Hardware Affects Provisioning Policies

- Hardware-driven refresh signals:
  - Newer is much better (H100 vs A100)→ Early decommission of old



Provision-Stage Findings: AI Hardware Affects Provisioning Policies

- Hardware-driven refresh signals:
  - Newer is much better (H100 vs A100)→ Early decommission of old

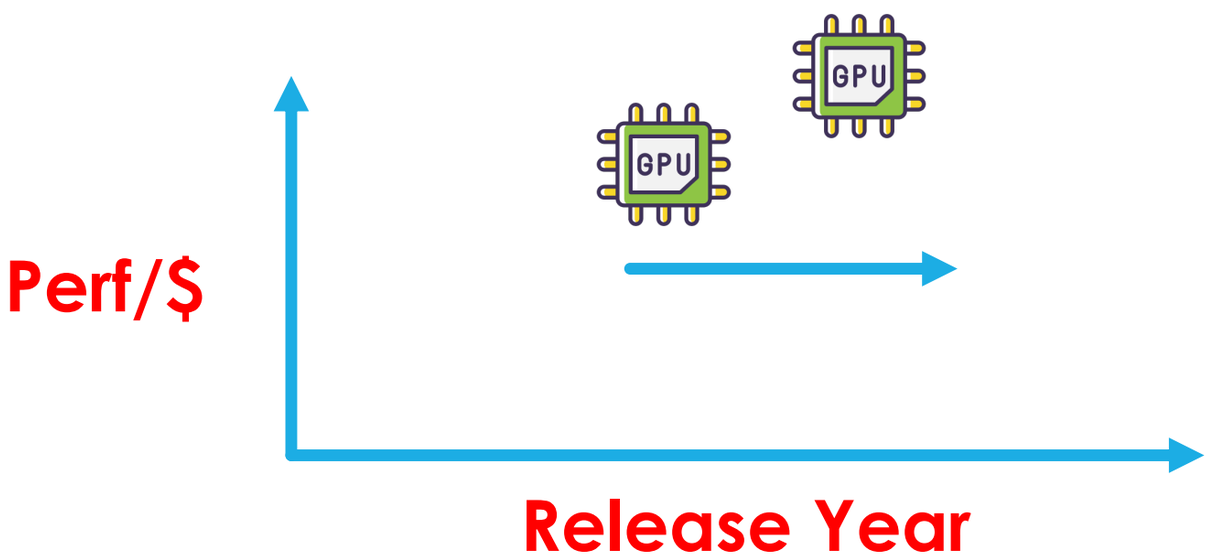


Provision-Stage Findings: AI Hardware Affects Provisioning Policies

- Hardware-driven refresh signals:
  - Newer is much better (H100 vs A100)→ Early decommission of old
  - Older is competitive (H100)

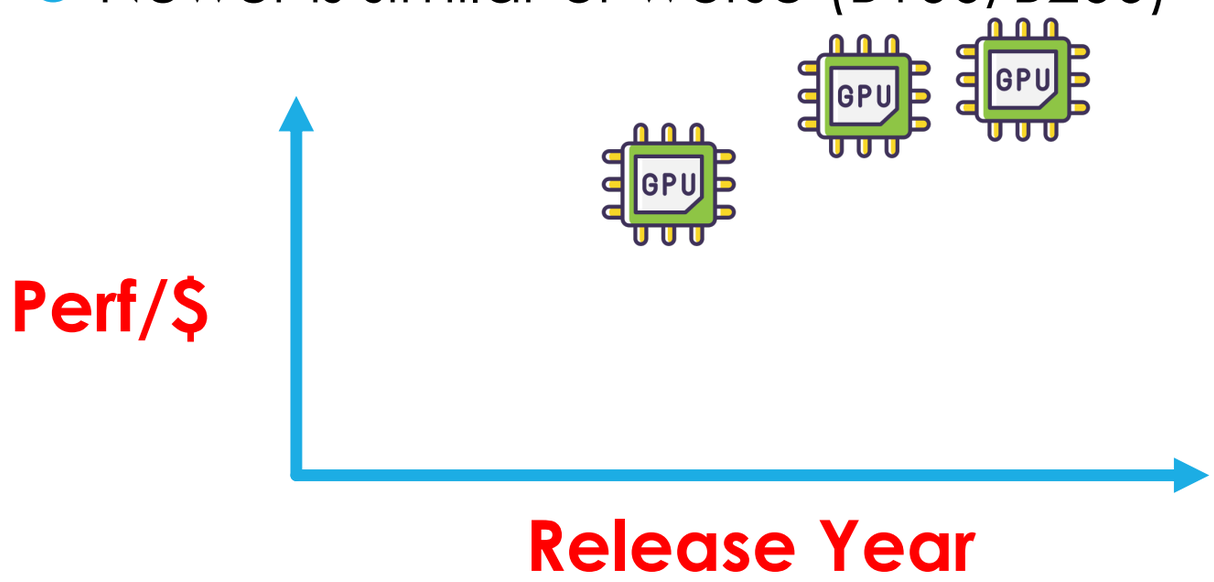
Provision-Stage Findings: AI Hardware Affects Provisioning Policies

- Hardware-driven refresh signals:
  - Newer is much better (H100 vs A100)→ Early decommission of old
  - Older is competitive (H100) → Extend lifetime



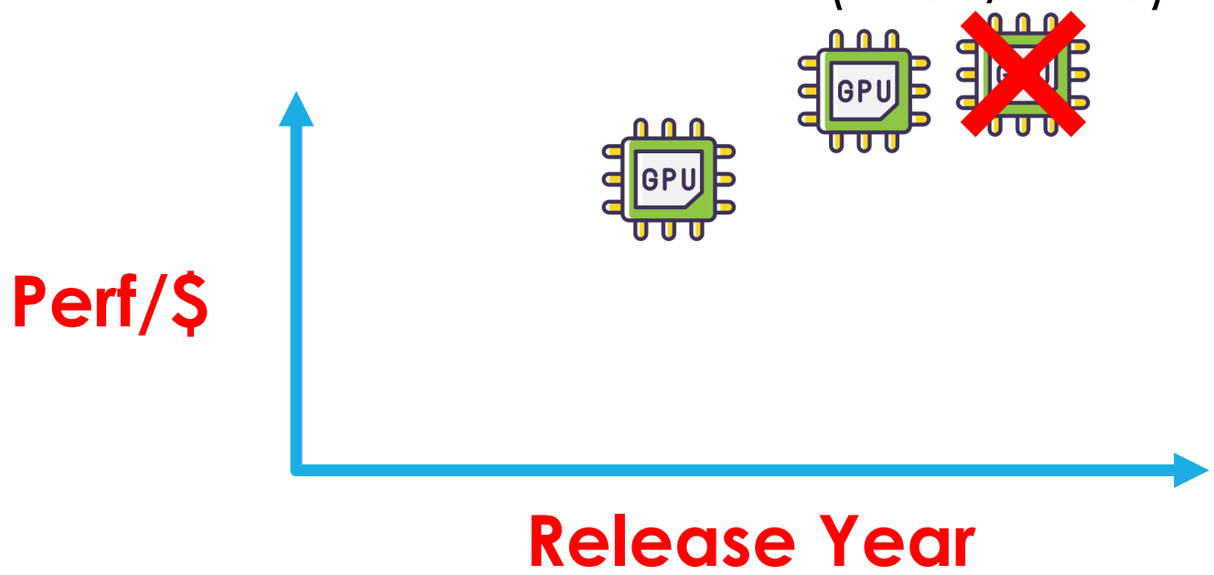
Provision-Stage Findings: AI Hardware Affects Provisioning Policies

- Hardware-driven refresh signals:
  - Newer is much better (H100 vs A100)→ Early decommission of old
  - Older is competitive (H100) → Extend lifetime
  - Newer is similar or worse (B100/B200)



Provision-Stage Findings: AI Hardware Affects Provisioning Policies

- Hardware-driven refresh signals:
  - Newer is much better (H100 vs A100)→ Early decommission of old
  - Older is competitive (H100) → Extend lifetime
  - Newer is similar or worse (B100/B200) → Skip generation

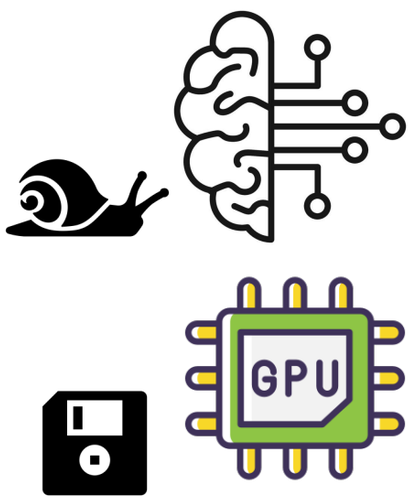


Provision-Stage Findings: AI Workload Affects Provisioning Policies

- Workload-driven refresh signals:

Provision-Stage Findings: AI Workload Affects Provisioning Policies

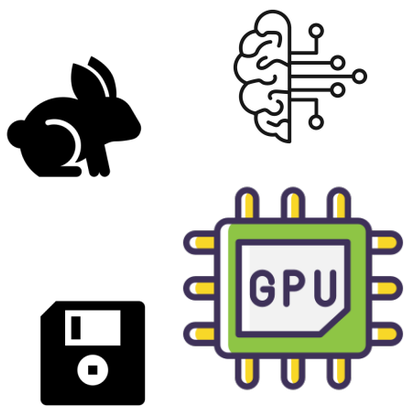
- Workload-driven refresh signals:
  - Smaller models remain efficient on older GPUs



Old GPU

Provision-Stage Findings: AI Workload Affects Provisioning Policies

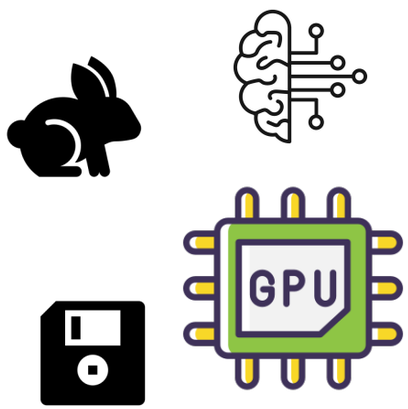
- Workload-driven refresh signals:
  - Smaller models remain efficient on older GPUs



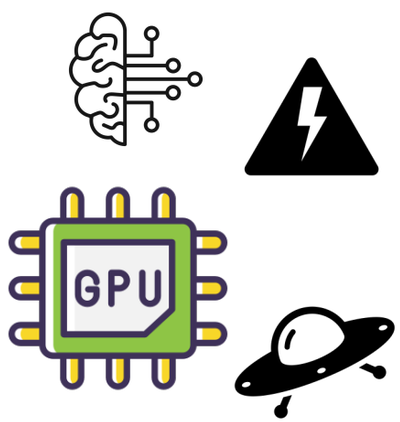
Old GPU

Provision-Stage Findings: AI Workload Affects Provisioning Policies

- Workload-driven refresh signals:
  - Smaller models remain efficient on older GPUs



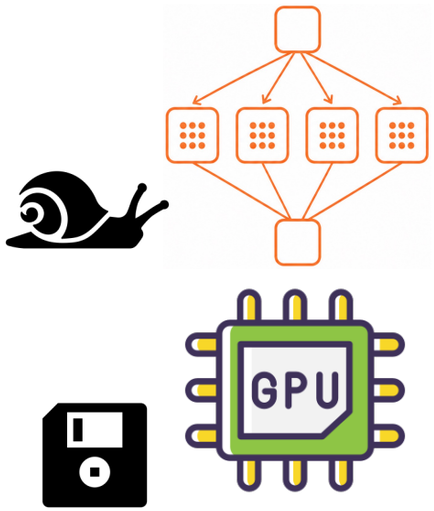
Old GPU



New GPU

Provision-Stage Findings: AI Workload Affects Provisioning Policies

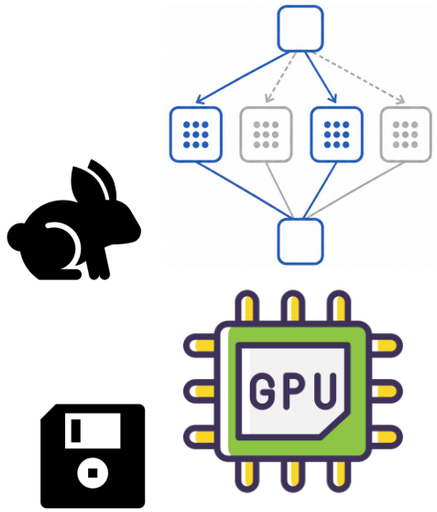
- Workload-driven refresh signals:
  - Sparse (MoE) models scale better on older GPUs



Old GPU

Provision-Stage Findings: AI Workload Affects Provisioning Policies

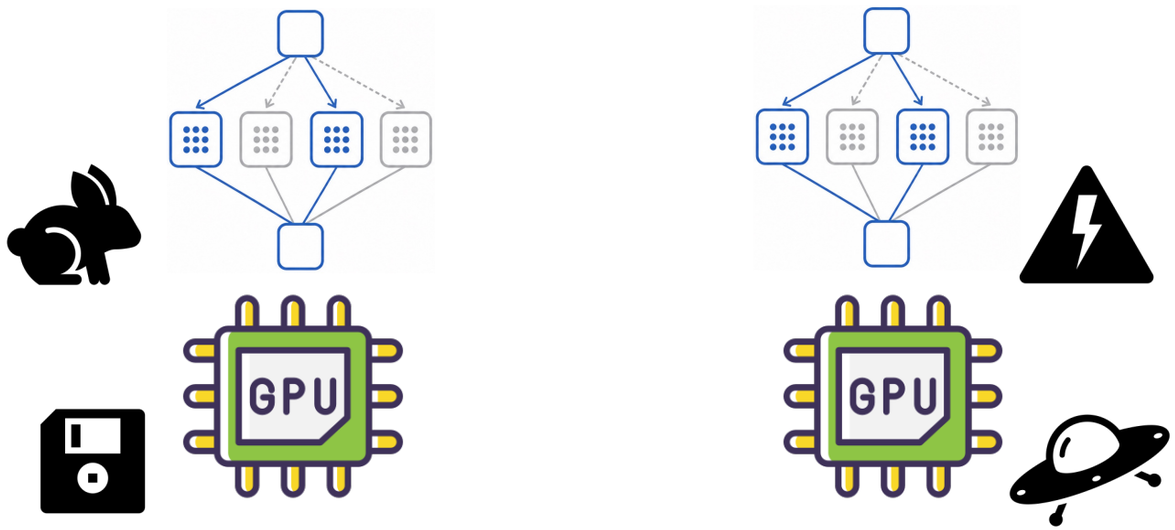
- Workload-driven refresh signals:
  - Sparse (MoE) models scale better on older GPUs



Old GPU

Provision-Stage Findings: AI Workload Affects Provisioning Policies

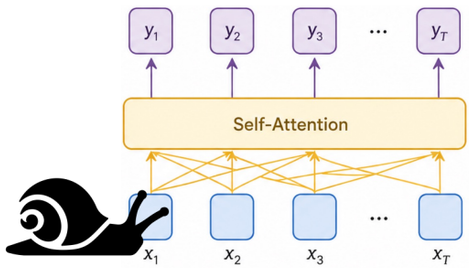
- Workload-driven refresh signals:
  - Sparse (MoE) models scale better on older GPUs

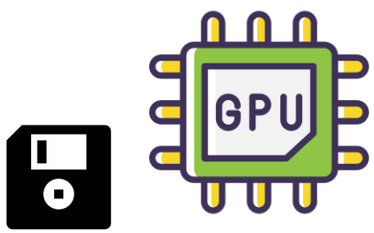


Old GPU **New GPU**

Provision-Stage Findings: AI Workload Affects Provisioning Policies

- Workload-driven refresh signals:
  - State-space models are more hardware-efficient for older GPUs

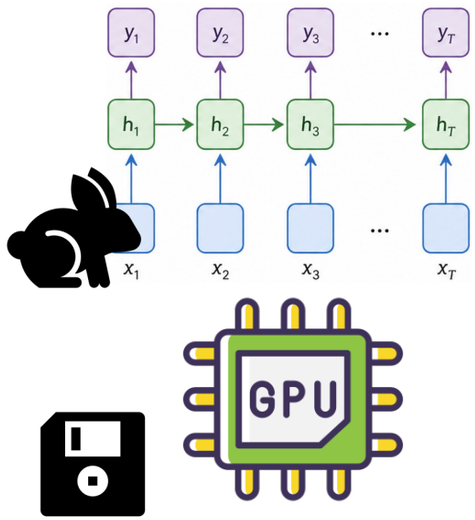




Old GPU

Provision-Stage Findings: AI Workload Affects Provisioning Policies

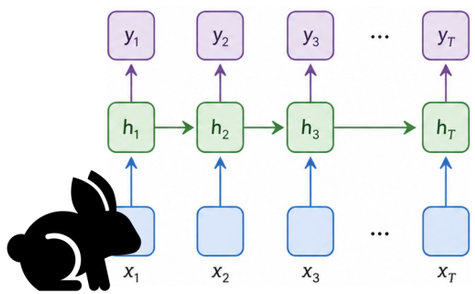
- Workload-driven refresh signals:
  - State-space models are more hardware-efficient for older GPUs

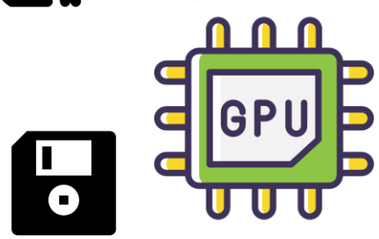


Old GPU

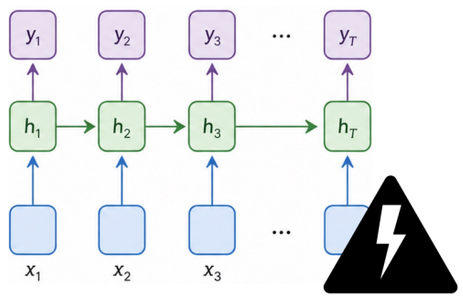
Provision-Stage Findings: AI Workload Affects Provisioning Policies

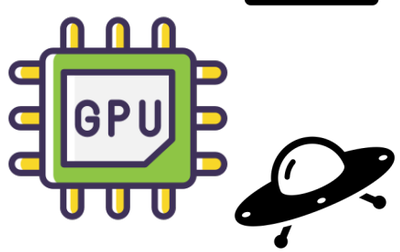
- Workload-driven refresh signals:
  - State-space models are more hardware-efficient for older GPUs





Old GPU

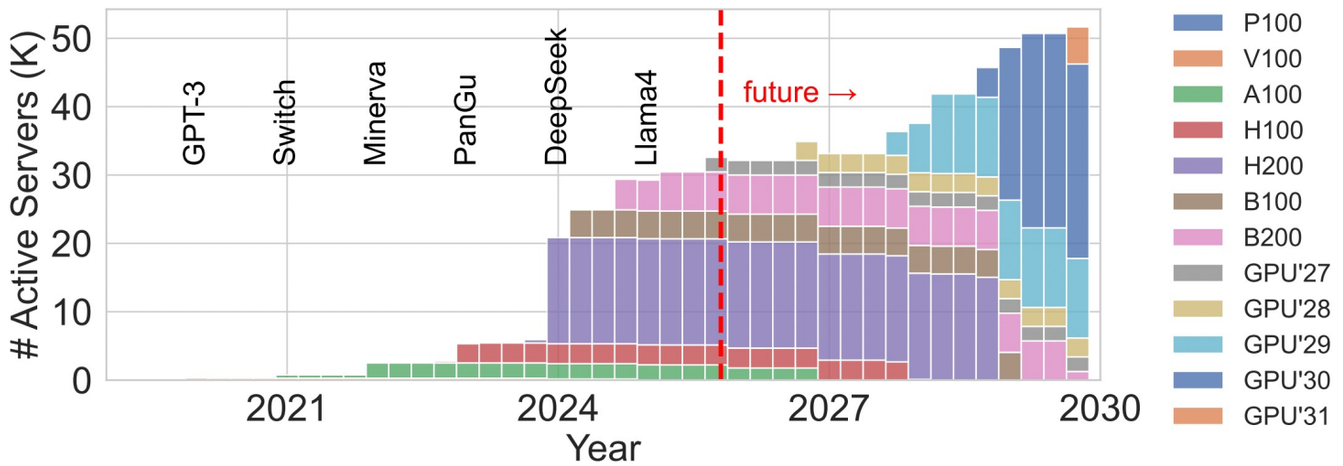




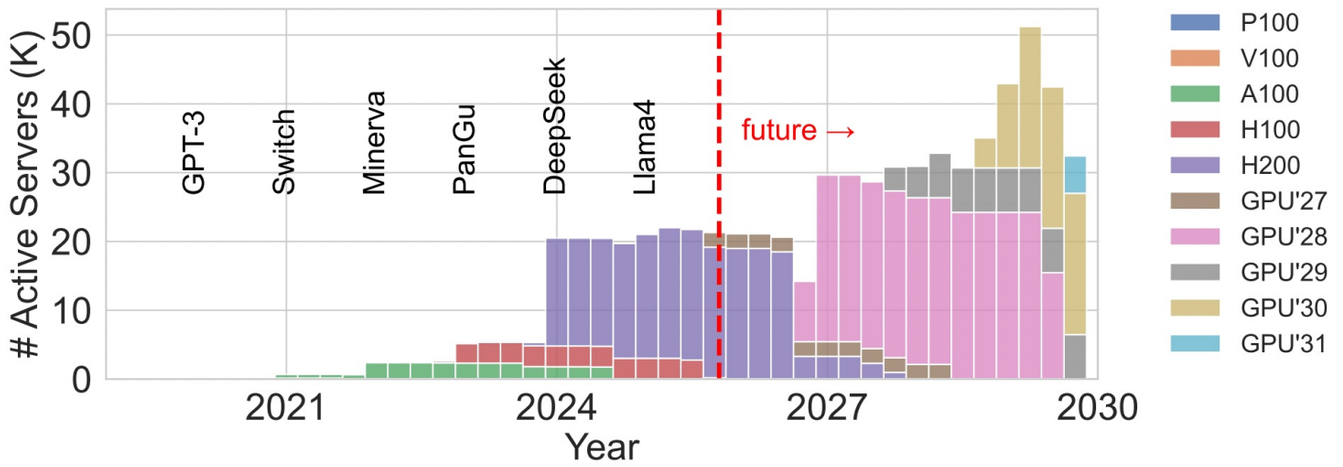
New GPU

Provision-Stage Findings: Need for Flexible IT Provisioning

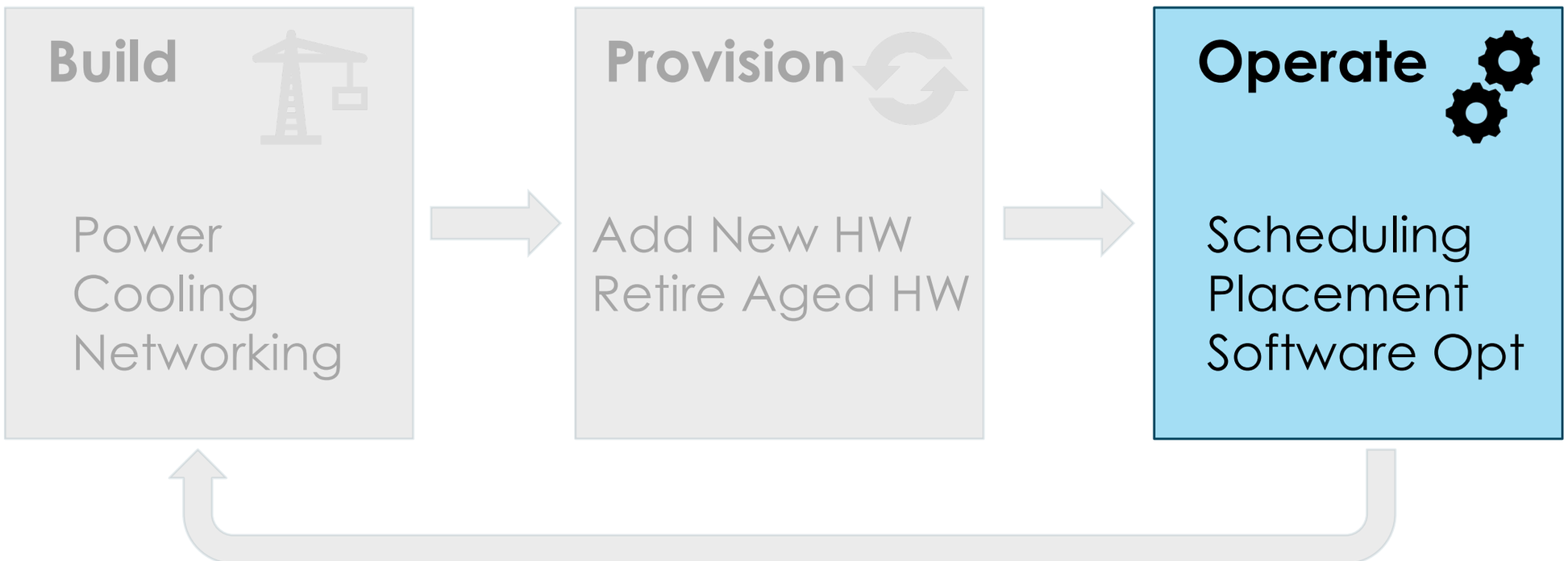
Traditional Approach



TCO-Optimal Approach



A Datacenter Lifecycle: Key Findings



Operate-Stage Findings: Homogeneous HW Pools in CPU Datacenters

- Traditional operation: workloads run on homogeneous HW pools with uniform deployment
- T-put and perf/Watt scale nearly linearly with newer servers

Operate-Stage Findings: Homogeneous HW Pools in CPU Datacenters

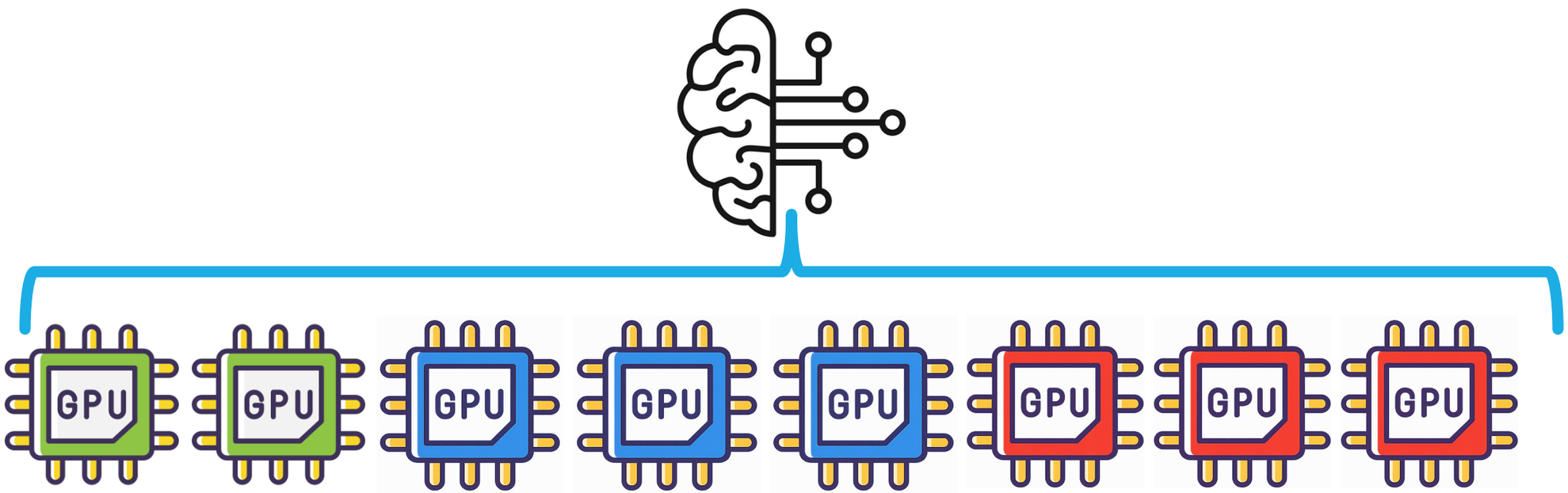
- Traditional operation: workloads run on homogeneous HW pools with uniform deployment
- T-put and perf/Watt scale nearly linearly with newer servers

Operate-Stage Findings: SW to Promote Heterogeneity in AI Datacenters

- AI workloads and hardware trends very different
- For cost-efficiency need to exploit heterogeneous fleet via SW

Operate-Stage Findings: SW to Promote Heterogeneity in AI Datacenters

- AI workloads and hardware trends very different
- For cost-efficiency need to exploit heterogeneous fleet via SW
- Quantization, P/D disaggregation, request routing…



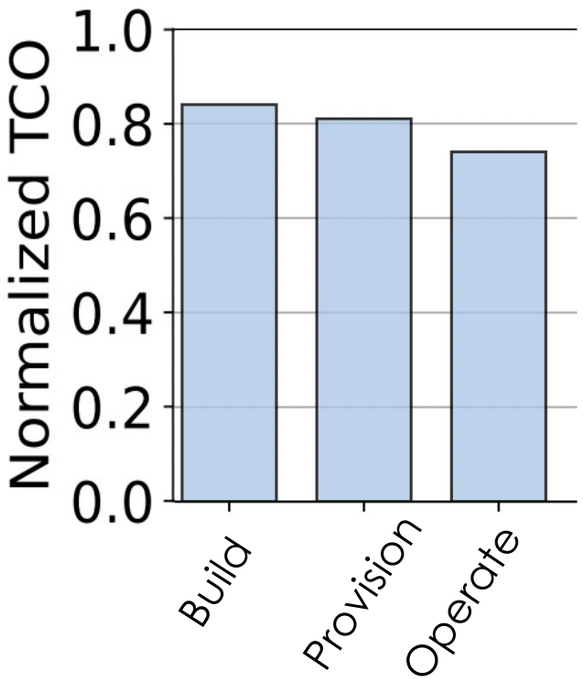
Operate-Stage Findings: SW to Promote Heterogeneity in AI Datacenters

- AI workloads and hardware trends very different
- For cost-efficiency need to exploit heterogeneous fleet via SW
- Quantization, …

Need for SW to align evolving models, growing demand, and a heterogeneous fleet

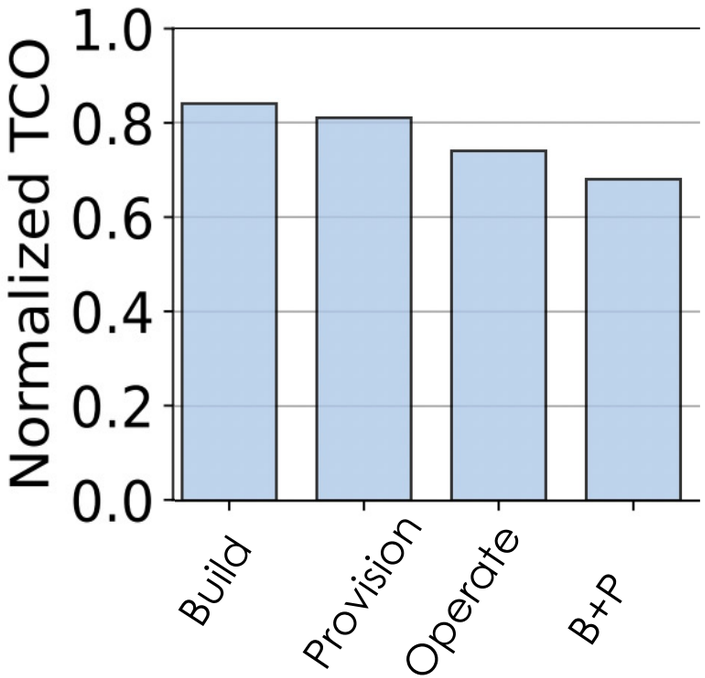
Cross-Stage Optimizations

- The largest TCO savings come from cross-stage rearchitecting the end-to-end AI datacenter lifecycle



Cross-Stage Optimizations

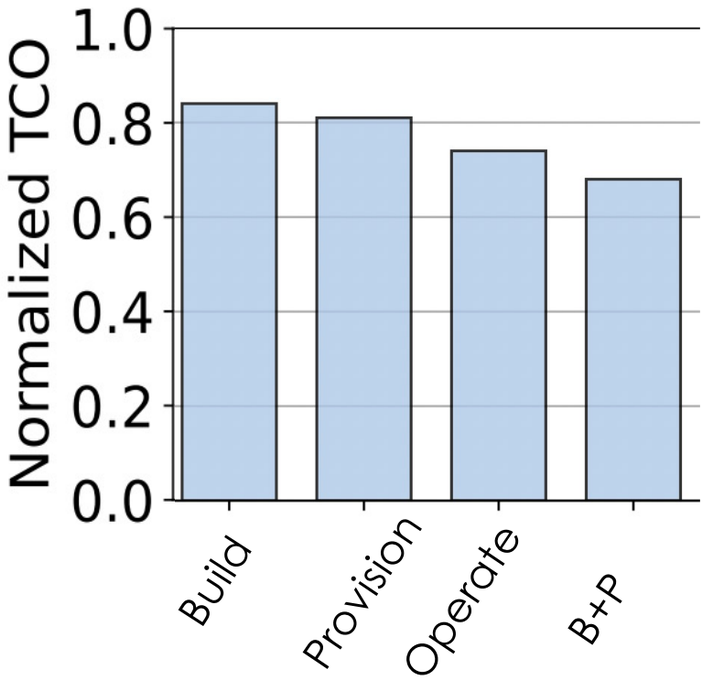
- The largest TCO savings come from cross-stage rearchitecting the end-to-end AI datacenter lifecycle



AI HW trends driving datacenter design: flat power, liquid cooling, NVLink

Cross-Stage Optimizations

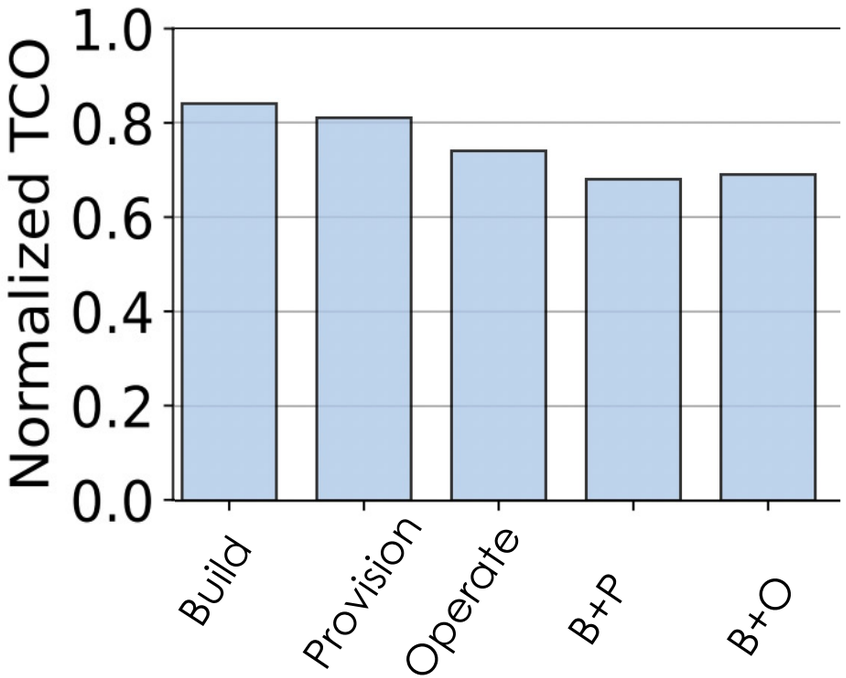
- The largest TCO savings come from cross-stage rearchitecting the end-to-end AI datacenter lifecycle



Higher upfront cost, but easier refreshes and longer lifetimes

Cross-Stage Optimizations

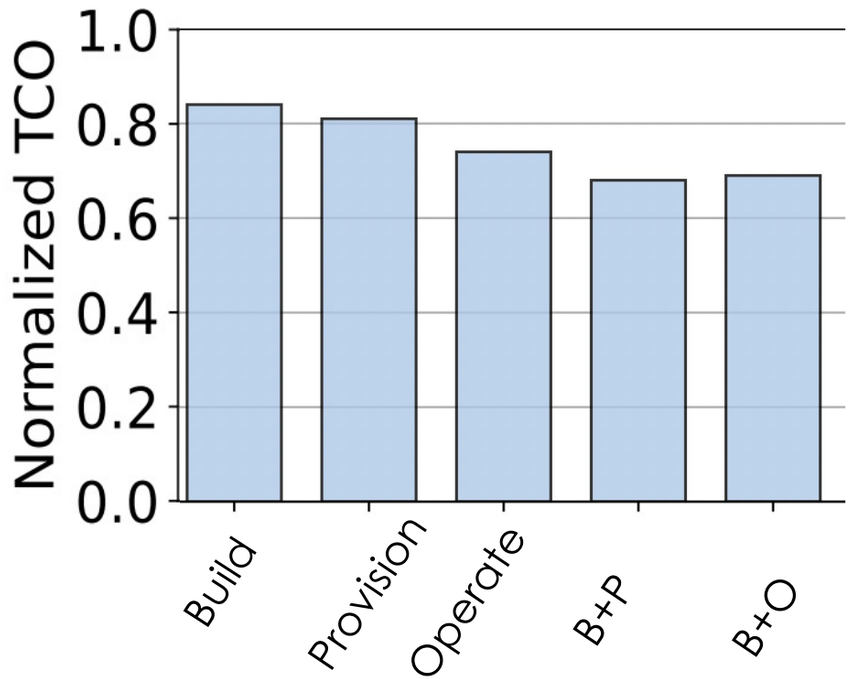
- The largest TCO savings come from cross-stage rearchitecting the end-to-end AI datacenter lifecycle



Infra decisions at build time shape operational flexibility

Cross-Stage Optimizations

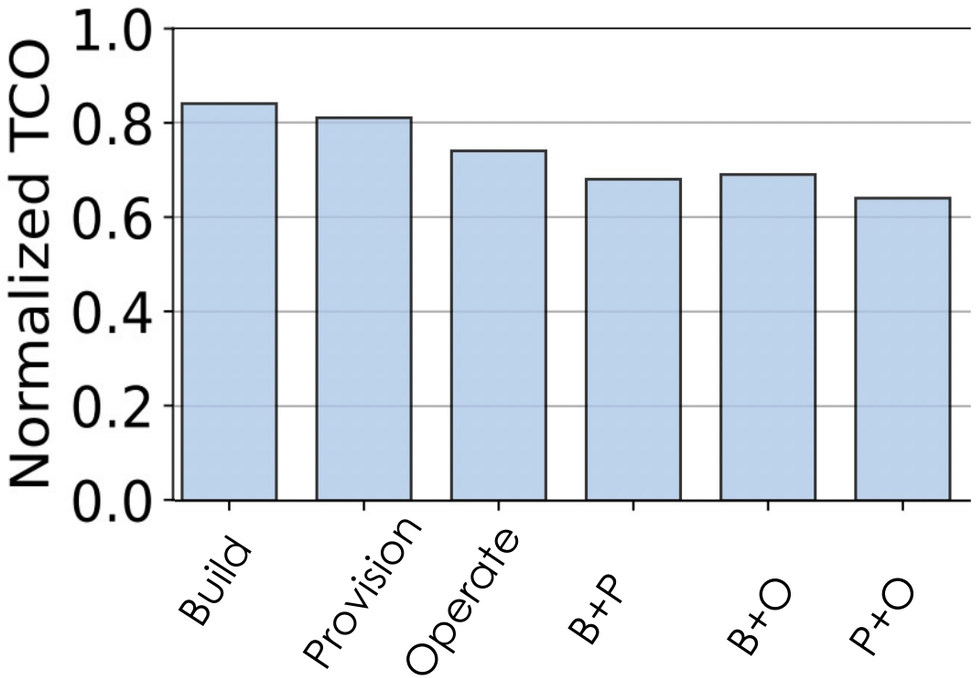
- The largest TCO savings come from cross-stage rearchitecting the end-to-end AI datacenter lifecycle



Runtime optimization boosts effective infrastructure capacity

Cross-Stage Optimizations

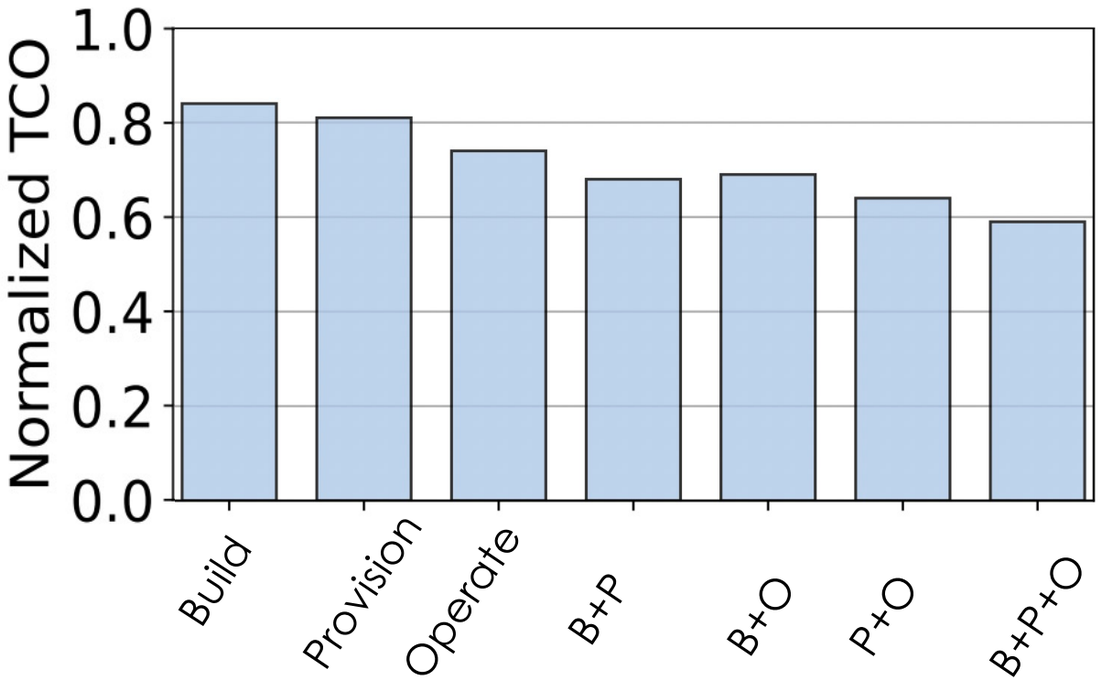
- The largest TCO savings come from cross-stage rearchitecting the end-to-end AI datacenter lifecycle



Heterogeneity-aware scheduling helps repurposing older GPUs

Cross-Stage Optimizations

- The largest TCO savings come from cross-stage rearchitecting the end-to-end AI datacenter lifecycle



Holistic approach reduces AI datacenter TCO by 40%!

More in the Paper

- Analysis of future trends in workloads and hardware
- Alternative future scenarios
  - Demand-shock, model size contraction, HW price/capability shock
- Sensitivity to trend changes

Conclusion

- Growth of AI has outpaced traditional datacenter management
- AI datacenters need flexible designs and lifecycle policies
- While lifecycle stage-specific improvements help, the biggest gains come from a holistic, cross-stage approach
  - Need to anticipate workload dynamics and hardware trends, enabling cloud providers to scale efficiently and control costs

Rearchitecting the Datacenter Lifecycle for AI

Open-Source Tool:

ISCA 2026

Jovan Stojkovic, Chaojie Zhang*, Íñigo Goiri*, Ricardo Bianchini*

The University of Texas at Austin, Microsoft Azure