# Disaggregating Persistent Memory and Controlling Them Remotely: An Exploration of Passive Disaggregated Key-Value Stores

*Shin-Yeh Tsai, Yizhou Shan, Yiying Zhang*

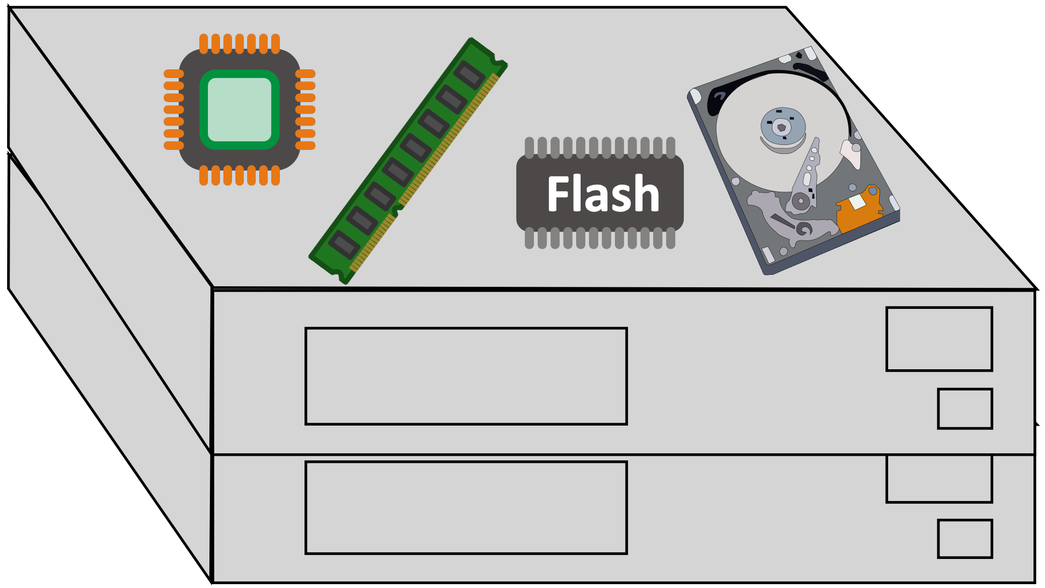
PURDUE UNIVERSITY

WukLab

UC San Diego

# Resource Disaggregation

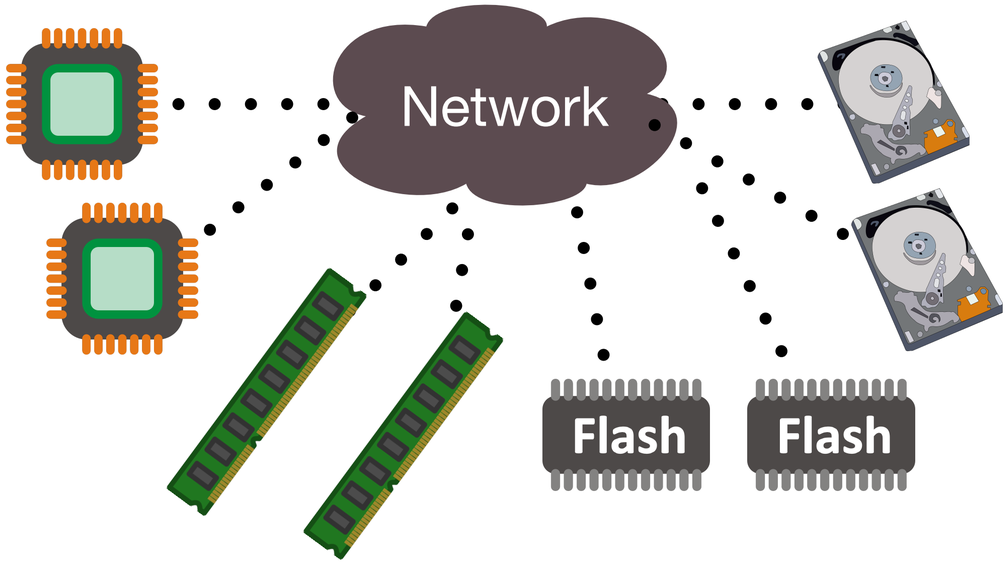
Break monolithic servers into ***network-attached*** resource pools



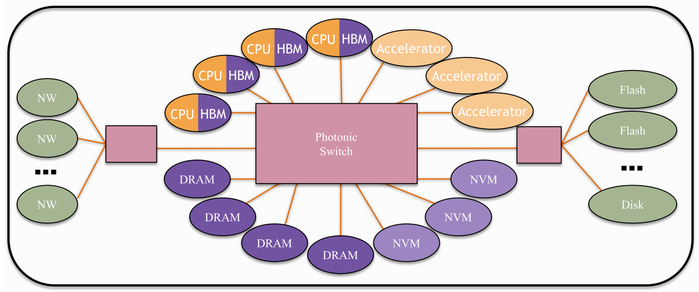
# Resource Disaggregation

Break monolithic servers into ***network-attached*** resource pools

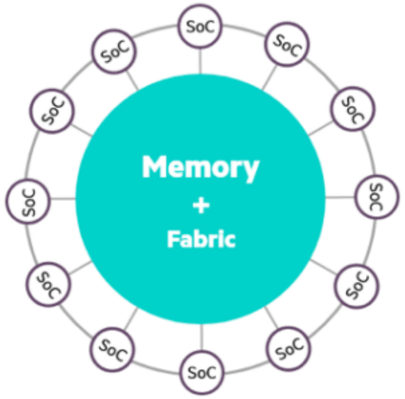
***Better manageability, independent scaling, tight resource packing***




***LegoOS***




Berkeley Firebox



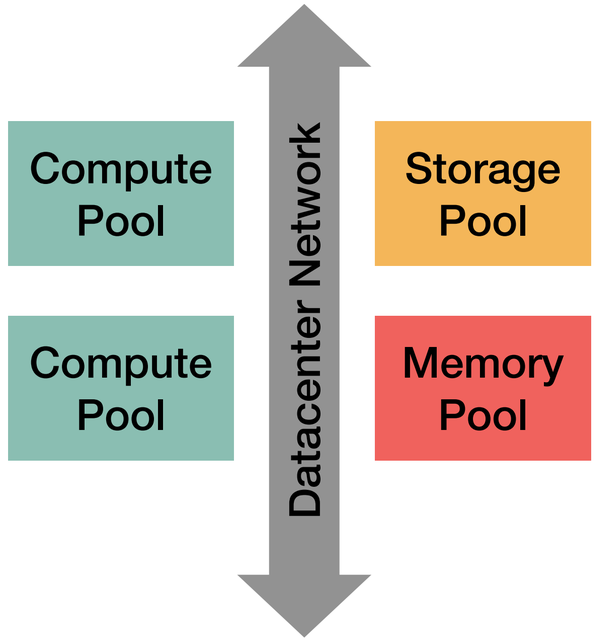

Hewlett Packard Enterprise

# Disaggregated Storage and Memory

Separate compute from storage and memory

- Manage and scale independently
- Apps access (large) non-local resources

A common practice in datacenters and clouds




amazon web services™ | S3

Powered By vmware vSAN™

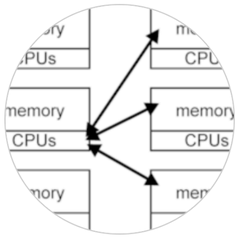
snowflake

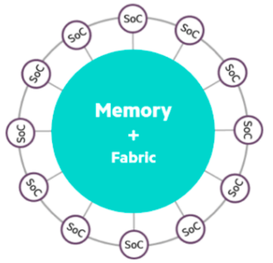
Microsoft

*FaRM*

vmware®

Remote memory




Hewlett Packard Enterprise

# Disaggregated Persistent Memory?

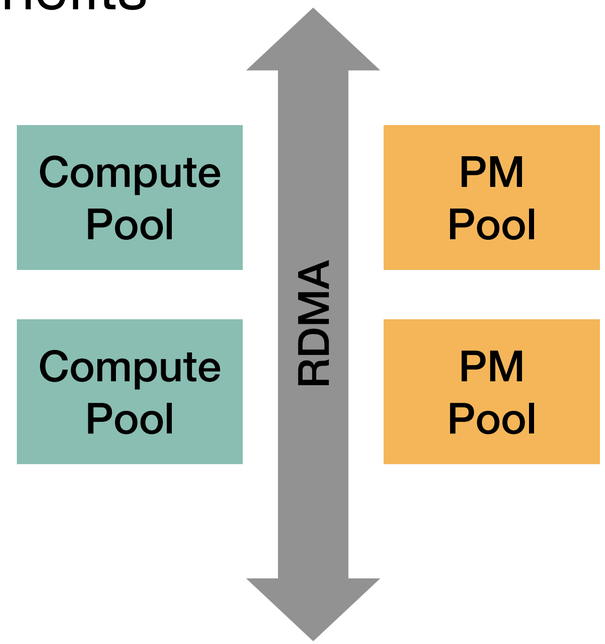
**PM:** byte-addressable, persistent, memory-like perf

**Disaggregating PM (DPM)**

- Enjoy disaggregation's management, scalability, utilization benefits
- Easy way to integrate PM into current datacenters

**Use existing disaggregated systems for DPM?**

- Disaggregated storage: software stack too slow for fast PM
- Disaggregated memory: do not provide data reliability

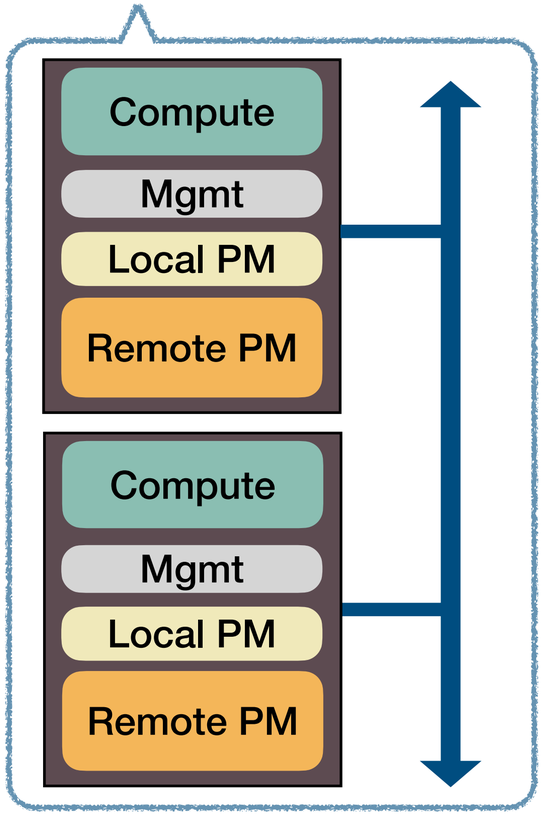


# Traditional Storage Systems

# Unexplored Area!

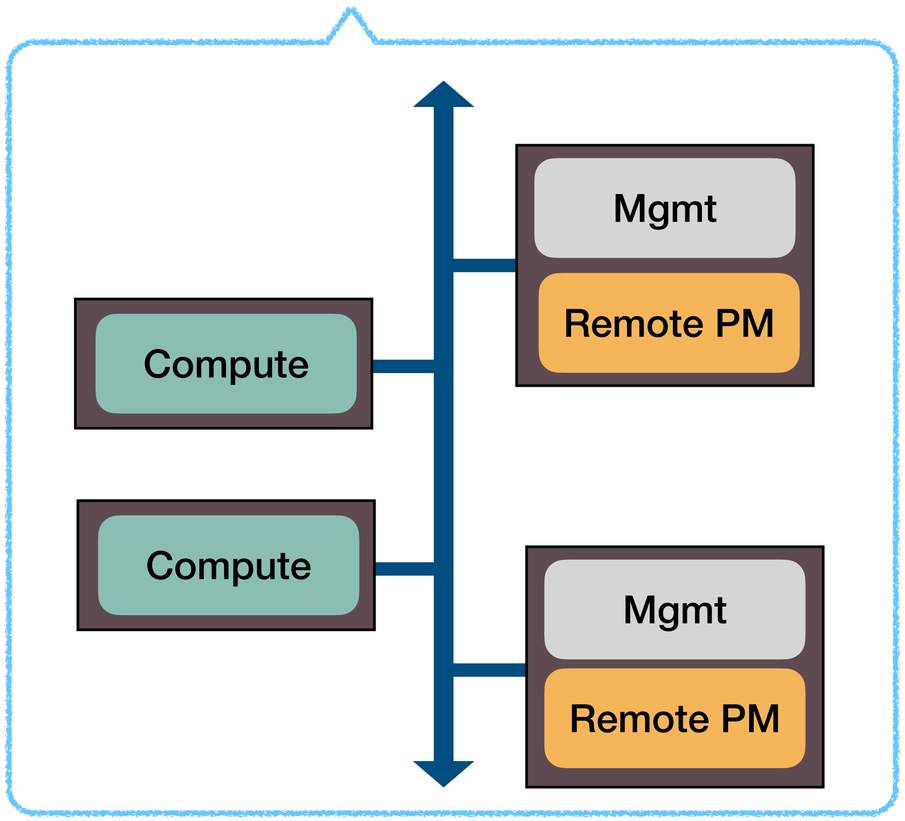
Spectrum of Datacenter PM Deploy Models

## Non-Disaggregation



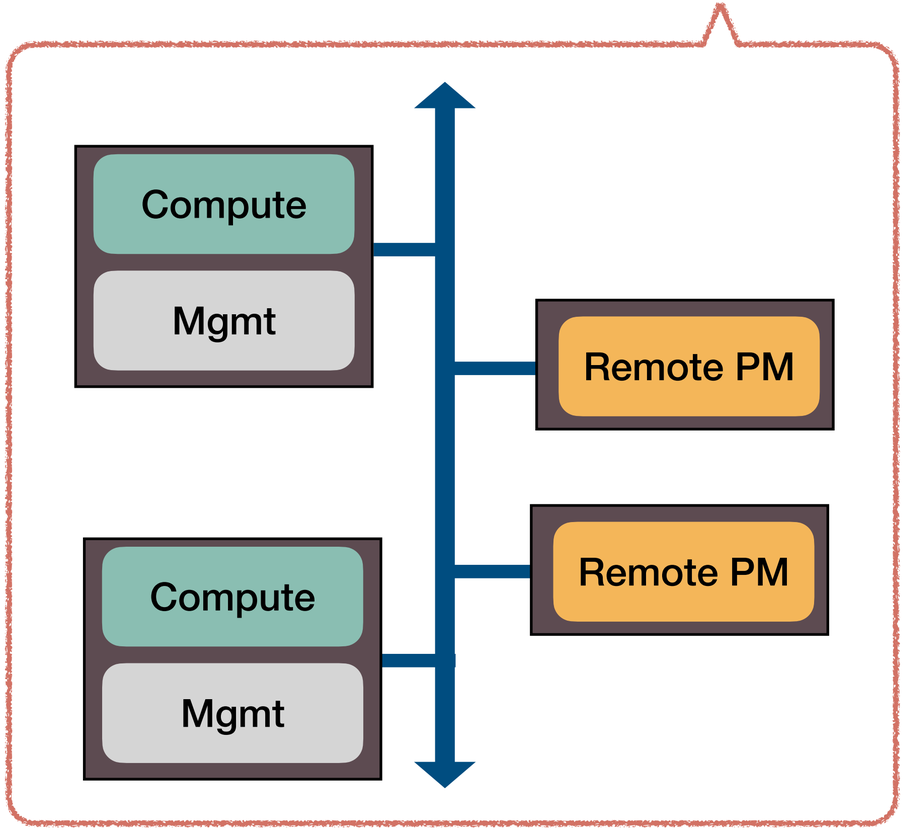

Hotpot [*SoCC'17*] (PM)
Octopus [*ATC'17*] (PM)
FaRM [*NSDI'14*] (DRAM)
Remote Regions [*ATC'18*] (DRAM)

## Active Disaggregation

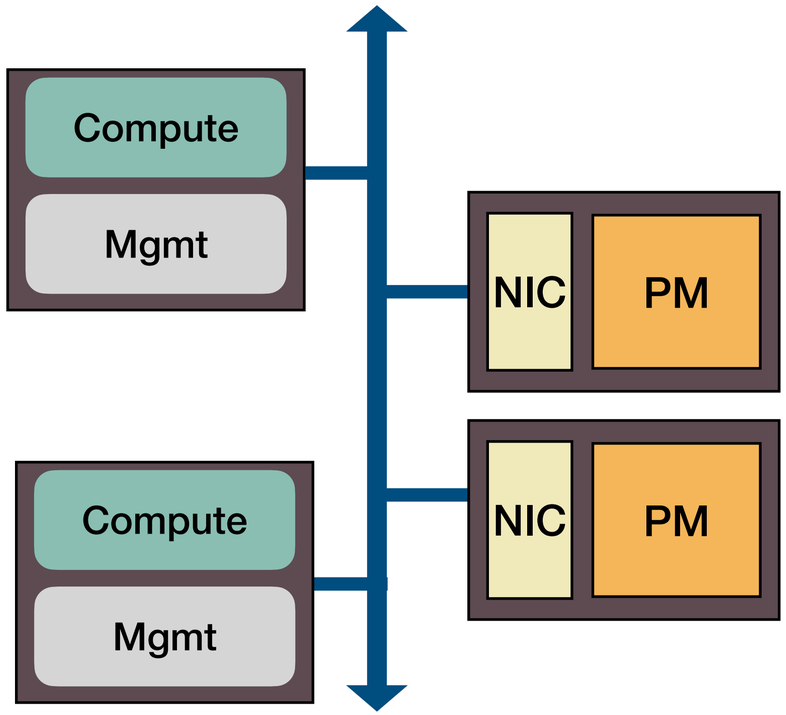



HERD [*SIGCOMM'14*] (in-DRAM KV store)
Decibel [*NSDI'17*] (cloud storage)
Snowflake [*NSDI'20*] (cloud storage)

## Passive Disaggregation



# Passive Disaggregated PM (pDPM)




**pDPM**

- Passive PM devices with NIC and PM
- Accessible only via network

**Why pDPM?**

- Low CapEx and OpEx
- Easy to add, remove, and change
- No scalability bottleneck at PM nodes
- Research value in exploring new design area

**Why possible now?** Fast RDMA network + CPU bypassing

***Without processing power at PM, where to process and manage data?***

Spectrum of Datacenter PM Deploy Models

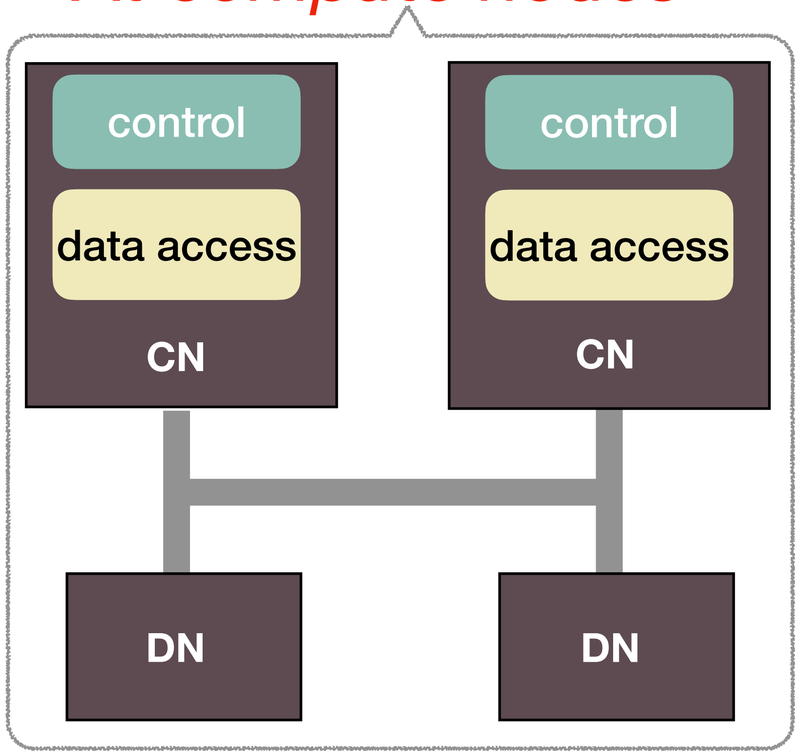
Non Disaggregation

Active Disaggregation

Passive Disaggregation

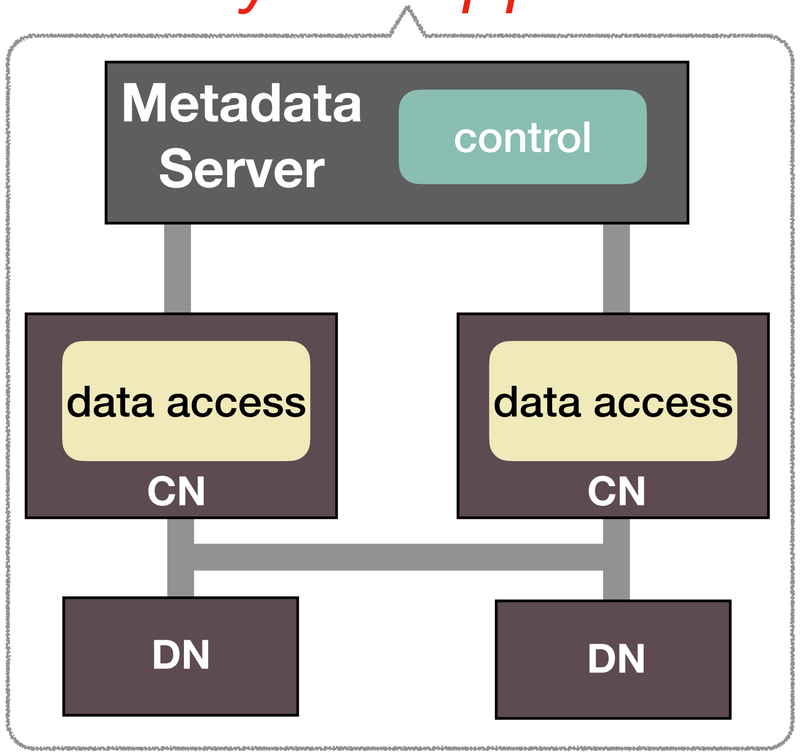
*Where to process and manage data?*

*At compute nodes*



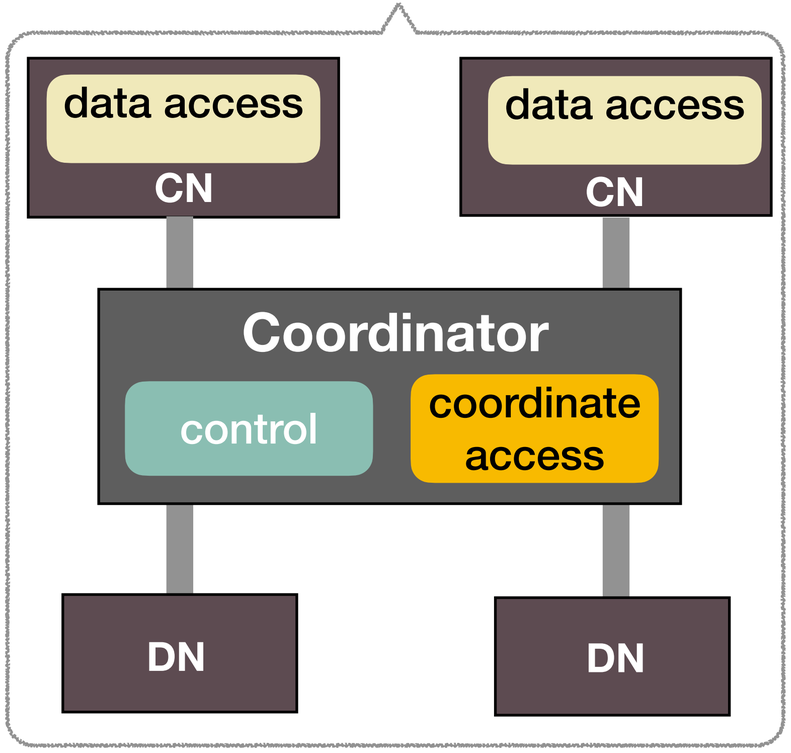

*Distributed data & metadata*

*A hybrid approach*




*Separate data & metadata*

*At a coordinator*




*Centralized data & metadata*

**CN**: Compute Node, **DN**: Data Node with PM

# Passive Disaggregated PM (pDPM) Systems

We design and implement three pDPM key-value stores

- At computer nodes → **pDPM-Direct**
- At global coordinator → **pDPM-Central**
- A hybrid approach → **Clover**

All guarantee ***read committed***, ***atomic write***, and ***data reliability***

# Results Highlight & Conclusion

- Clover is the best pDPM model

=> Clean separation of data and metadata planes is the key!

- pDPM can achieve similar performance as active DPM, but at a lower cost
- pDPM performs worse under heavy write contention
- Future system could benefit from a hybrid disaggregation model

# Thank you!

open source @ github.com/WukLab/pDPM

WukLab.io

PURDUE UNIVERSITY

UC San Diego