AgileMem

DRAM Memory

- DRAM consumes a lot of power when active-idle

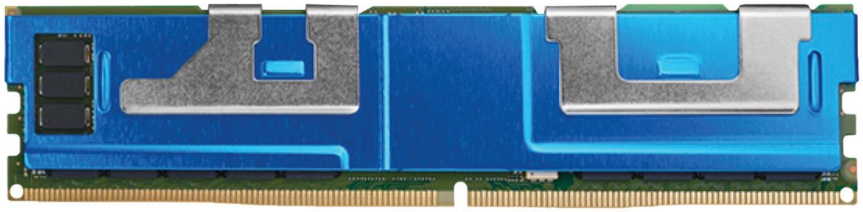
DRAM DIMM

- DRAM power states reduce idle power consumption but may hurt performance

DRAM Power State	Description	Transition Latency	PkgC-State Prerequisite	Power (16GB Device)
Active idle				~1.6W
Shallow idle (CKE-OFF)	DDR clock is disabled	10 - 30 ns	PC1A	0.43 - 1.01W
Deep idle (Self-Refresh)	DRAM is responsible for the refresh process	5 - 7 us	PC6	~0.25W

Optane DC Persistent Memory (PMM)

- Offers fast (< 1us) persistent storage



Optane DIMM

- Uses DIMM slots like DRAM
- Power consumption
  - Active power: Consumes more power than DRAM due to underlying technology
  - Idle power: Unknown if Optane PMM has power states
- Discontinued in July 2022
  - But other similar technologies are emerging e.g. Weebit ReRAM

Our Goal

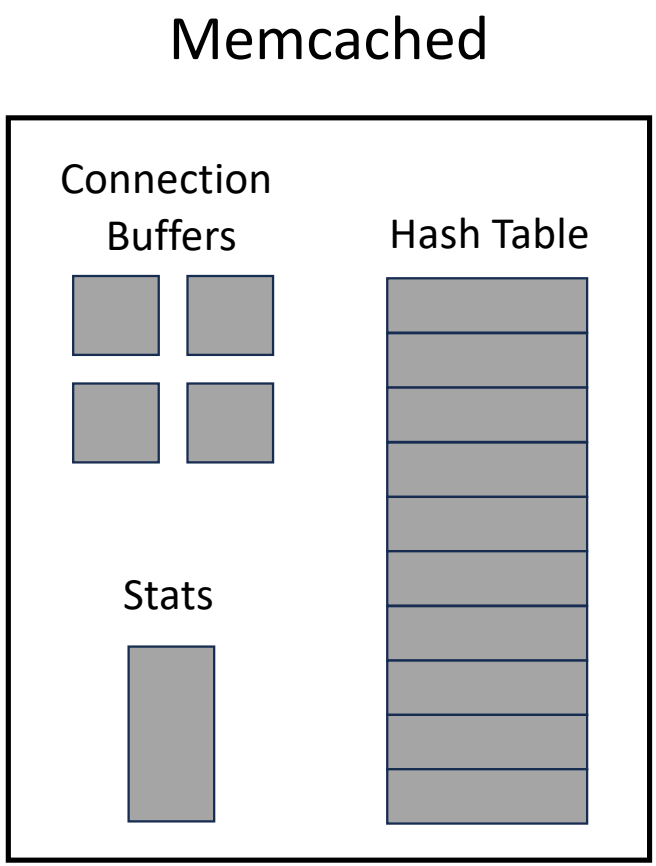
Leverage Optane to reduce idle power

- Characterize Optane power consumption
- Use Optane to reduce memory power consumption at low utilization

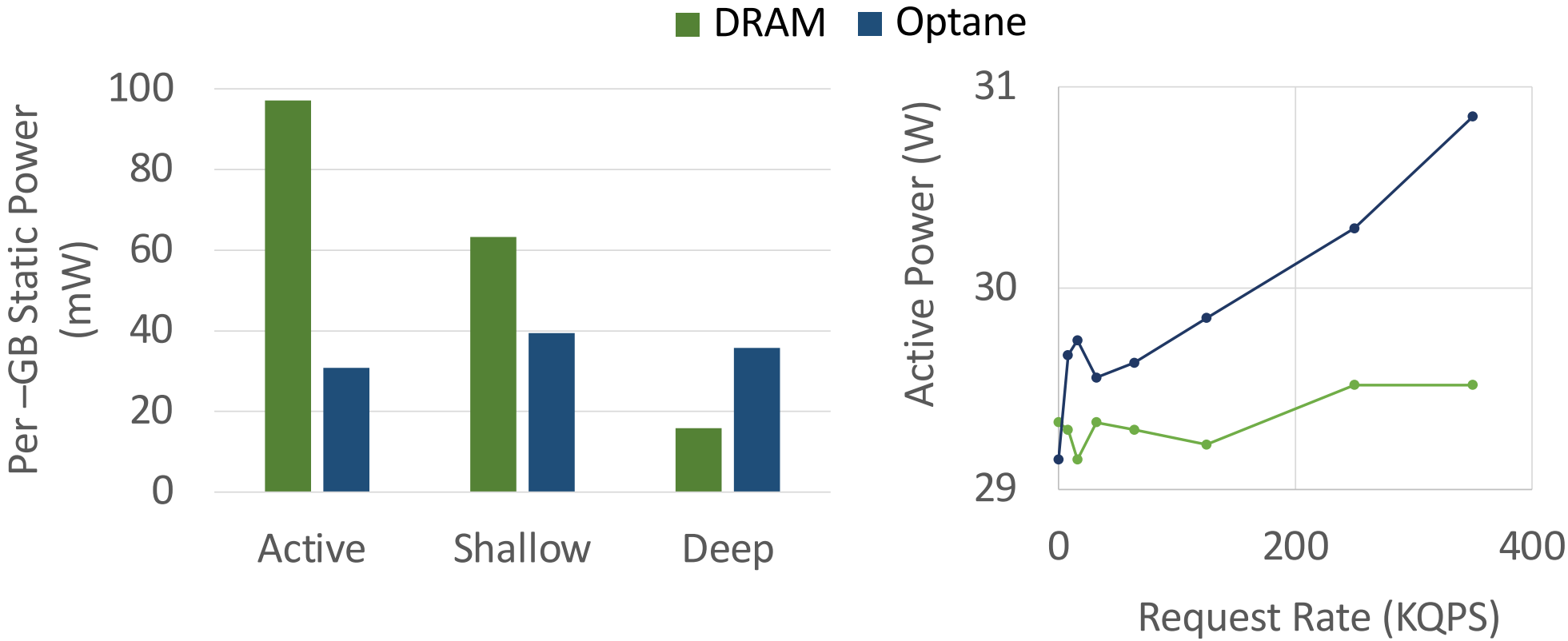
Characterizing Optane impact on Memcached

Experiment Configurations:

- DRAM-only
  - Uses standard `malloc` for memory allocation in DRAM
  - Key points: low latency, limited capacity
- Optane-only
  - Uses **`libvmmalloc`** to redirect memory allocation to Optane
  - Key points: larger capacity, higher latency



DRAM vs Optane

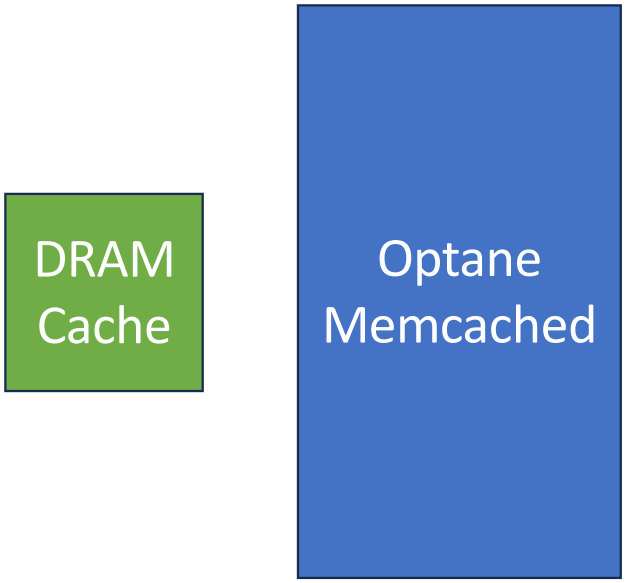


- Optane idle power is unexpectedly high, with no power saving states
- Optane power increases with workload activity

Leveraging Optane to reduce idle power

Leveraging Optane to reduce idle power

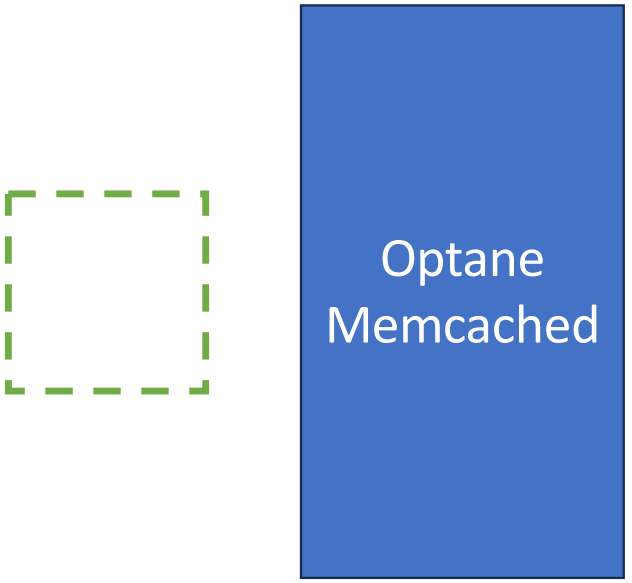
- Cache frequently accessed data
  in DRAM using write-through caching



Leveraging Optane to reduce idle power

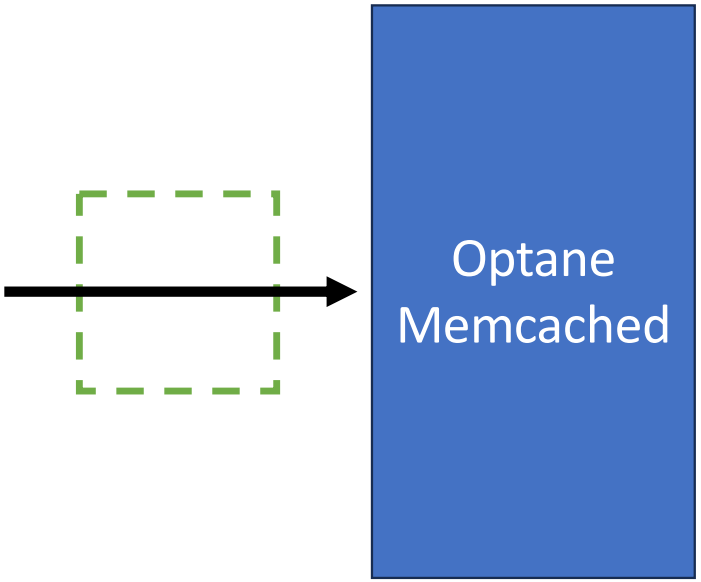
- Cache frequently accessed data in DRAM using write-through caching

- Allow DRAM to enter self-refresh while system is idle

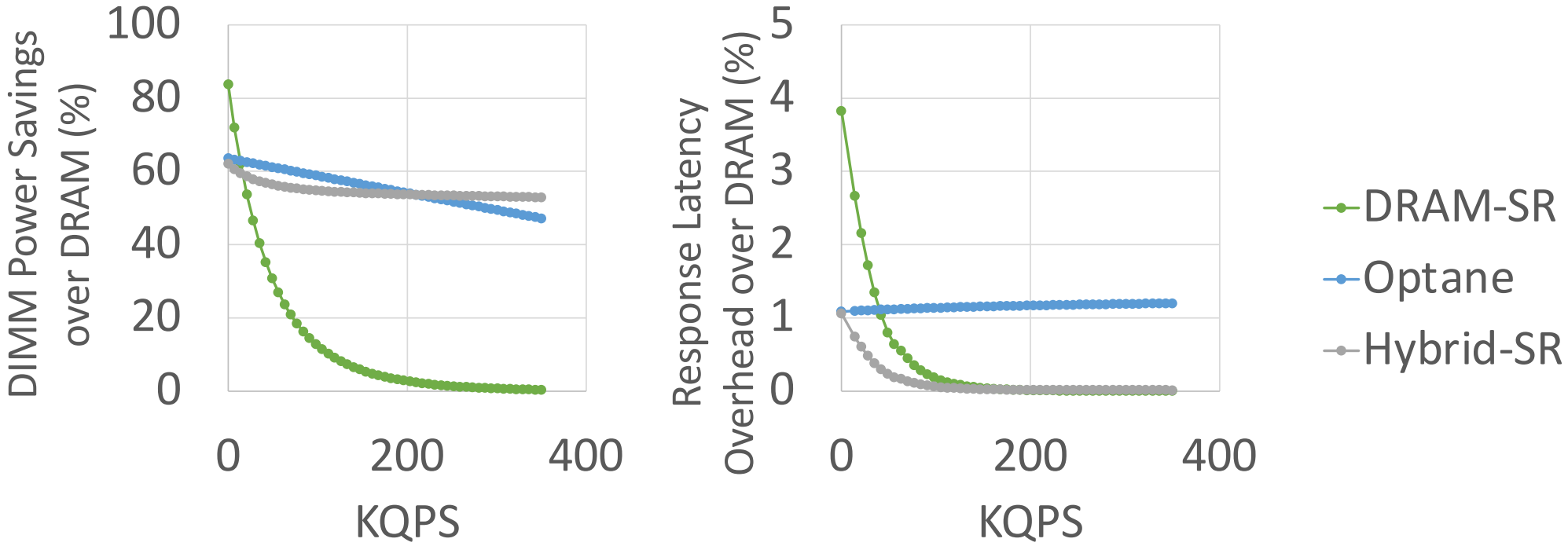


Leveraging Optane to reduce idle power

- Cache frequently accessed data in DRAM using write-through caching
- Allow DRAM to enter self-refresh while system is idle
- Serve queries from Optane while DRAM is waking up from self-refresh



Memcached (modeled, isocapacity)



- DRAM-SR: Self-refresh reduces power but increases latency at low QPS
- Optane: Avoids self-refresh latency but increases latency at high QPS
- Hybrid-SR: Provides low latency, and reduces Optane dynamic power

Summary

- DRAM self-refresh enables significant power savings but hurts latency
- Optane has low idle power without power states
- Combination shows potential for both reducing power and maintaining performance