

KV Cache is a Cost Question

Jim Handy

OBJECTIVE ANALYSIS

Outline

- Why have a KV Cache?
- What Kind of KV Cache is Best?
- Where Does KV Cache Go?
- Tough KV Cache Questions
- Parting Thoughts

How to Think of a KV Cache

- Remember website FAQs?
  - Website FAQs provided answers to common questions
    - The idea was to reduce the load on employees
  - Questions & answers weren't well vetted
    - No statistical basis for "Frequently Asked"
  - KV Cache is improved FAQs for inference























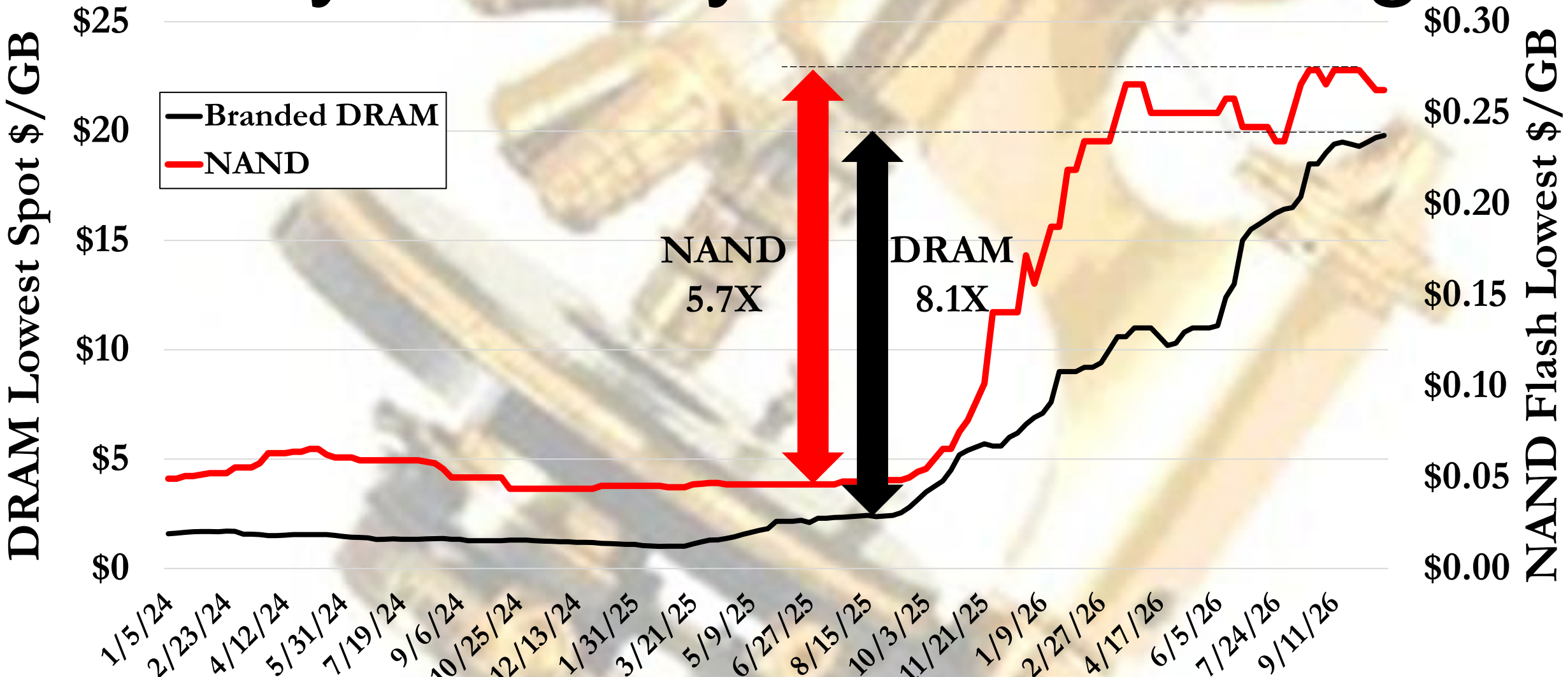



STORE
CALCULATE
Extracted Features
KV Cache
Fish
Kind of Fish?
Not Fish
Do Complete Identification

KV Cache: A Financial Balance

- Calculate or Store?
  - Depends on the cost of calculation vs. storage
    - Frequently-used information is stored
    - Infrequently-used information is calculated every time
  - Memory and storage cost a lot these days
    - Balance changes in proportion to cost of storage

Today Memory Prices are High



Where to Keep the KV Cache?

- HBM
- System memory
- Remote or pooled memory
- Storage

- $$$$$$$!!!
- Cheaper, but small
- CXL, possibly re-using DDR4
- SSD/HDD

It's About Time vs. Money

CXL Market Now Taking Shape



- New report from Objective Analysis
- Covers all perspectives
  - Where CXL is useful, and where it isn't
  - Demand drivers for CXL DRAM modules
  - Opportunities outside of DRAM
  - Forecast (Revenues, Units, ASP)
- Available for immediate download:
  - Objective-Analysis.com/reports

Tough KV Cache Questions

- What goes in, what stays out?
- Where to put it: HBM, Memory, CXL, SSD?
  - Split it across tiers?
- Maximum size?
- What if memory prices fall (or rise)?
- Effects of a model change

There's a Lot to Think About

- Best KV Cache size
- Best KV Cache location
- Managing cache data
- Planning for future models
- What to do when prices go up or down?