

Learning to Scale Out By Scaling Down

The FAWN Project

David Andersen, Vijay Vasudevan, Michael Kaminsky*, Michael A. Kozuch*, Amar Phanishayee, Lawrence Tan, Jason Franklin, Iulian Moraru, Sang Kil Cha, Hyeontaek Lim, Bin Fan, Reinhard Munz, Nathan Wan, Jack Ferris, Hrishikesh Amur**, Wolfgang Richter, Michael Freedman***, Wyatt Lloyd***, Padmanabhan Pillali*, Dong Zhou

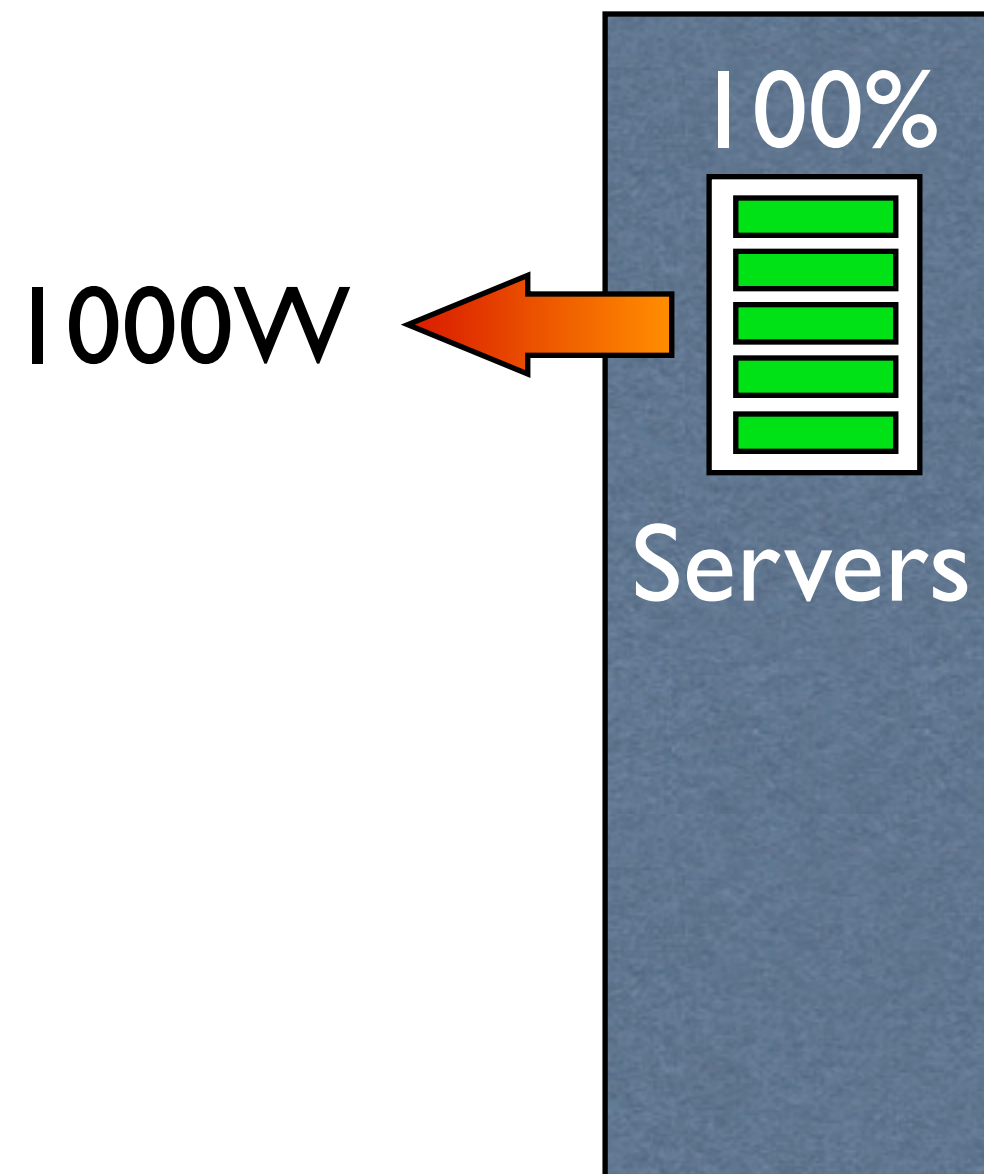
Carnegie Mellon University

** Princeton University

*Intel Labs Pittsburgh

*** Georgia Tech



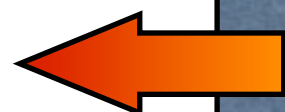




2000W



1000W



100%



Servers

Infrastructure: PUE

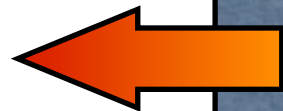
2005: 2–3

2012: ~1.1

Leave it to industry



1000W



100%



Servers

Infrastructure: PUE

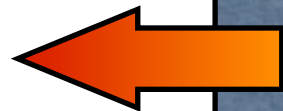
2005: 2–3

2012: ~1.1

Leave it to industry



1000W



100%



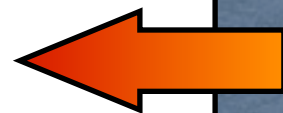
Servers

20%



Proportionality

750W



200W

Infrastructure: PUE

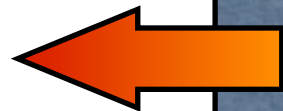
2005: 2–3

2012: ~1.1

Leave it to industry



1000W

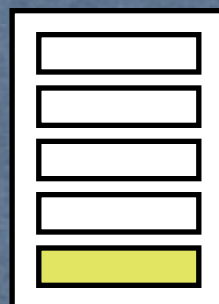


100%



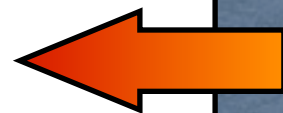
Servers

20%



Proportionality

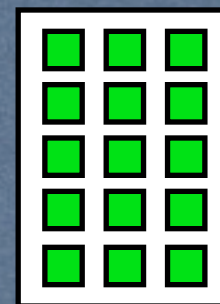
750W



200W

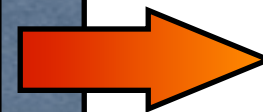


100%



FAWNs

Efficiency



300W

Infrastructure: PUE

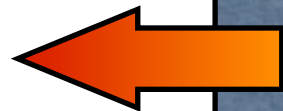
2005: 2–3

2012: ~1.1

Leave it to industry



1000W



100%



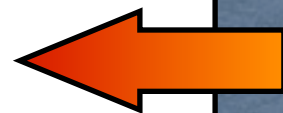
Servers

20%

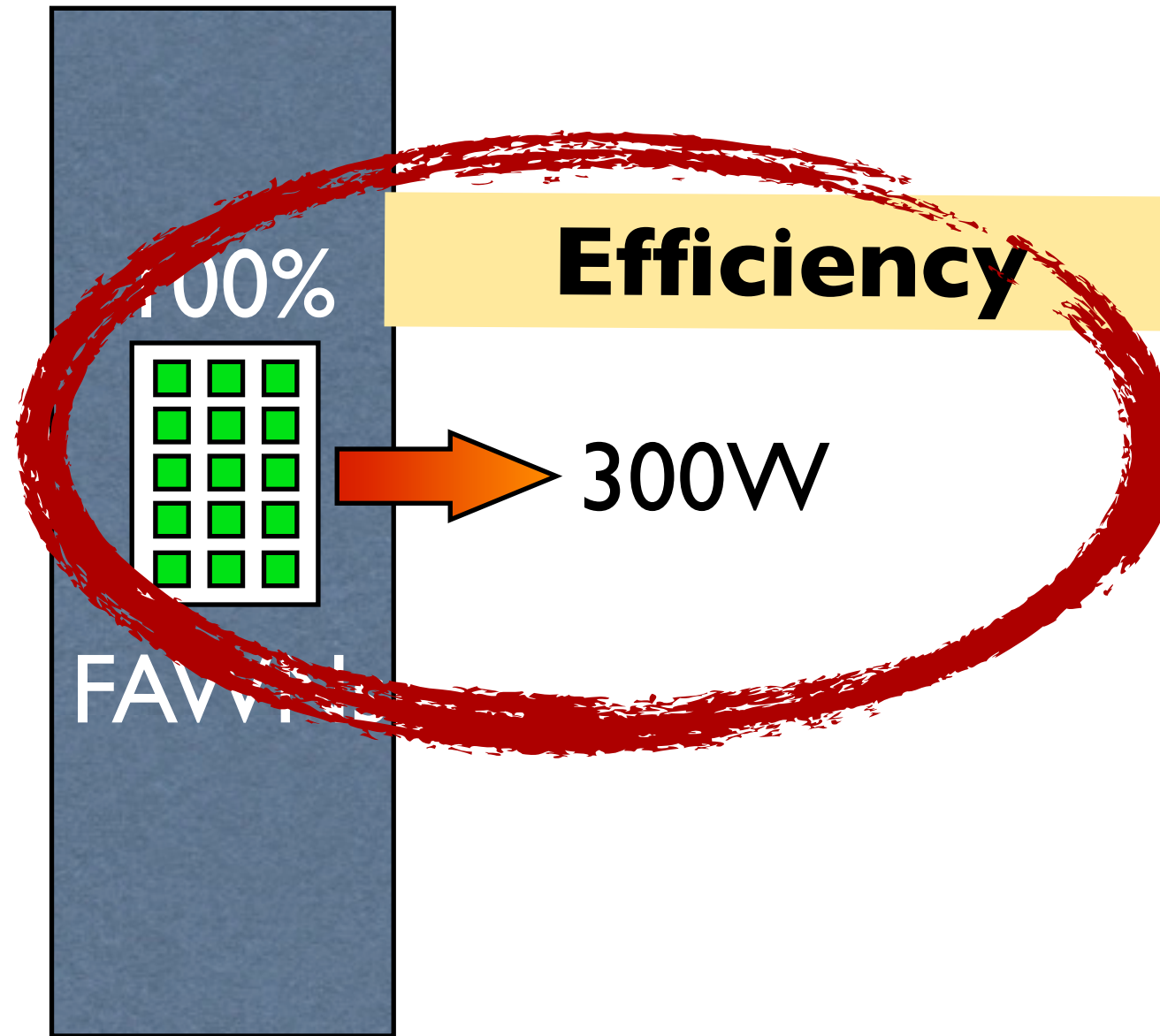


Proportionality

750W

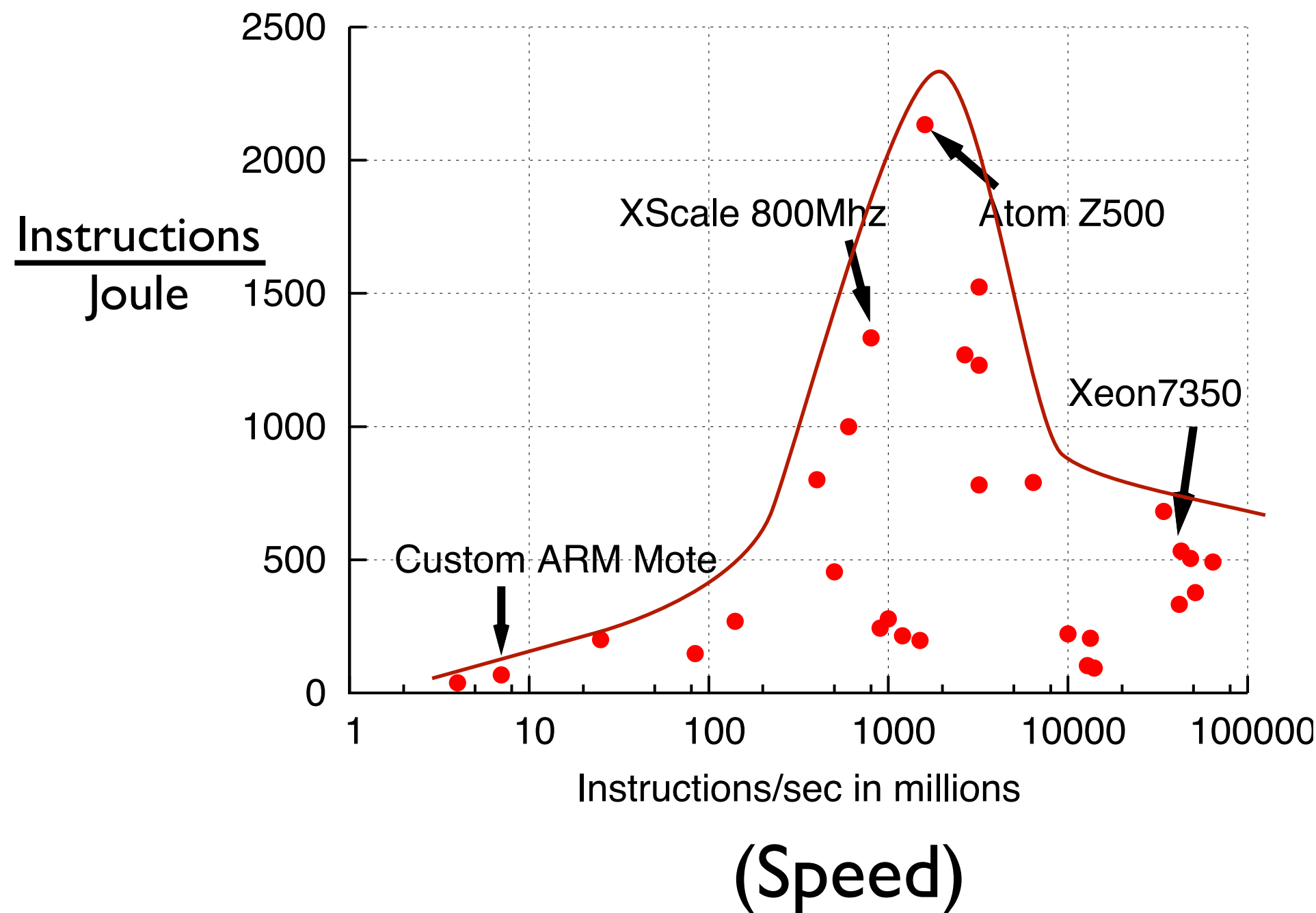


200W

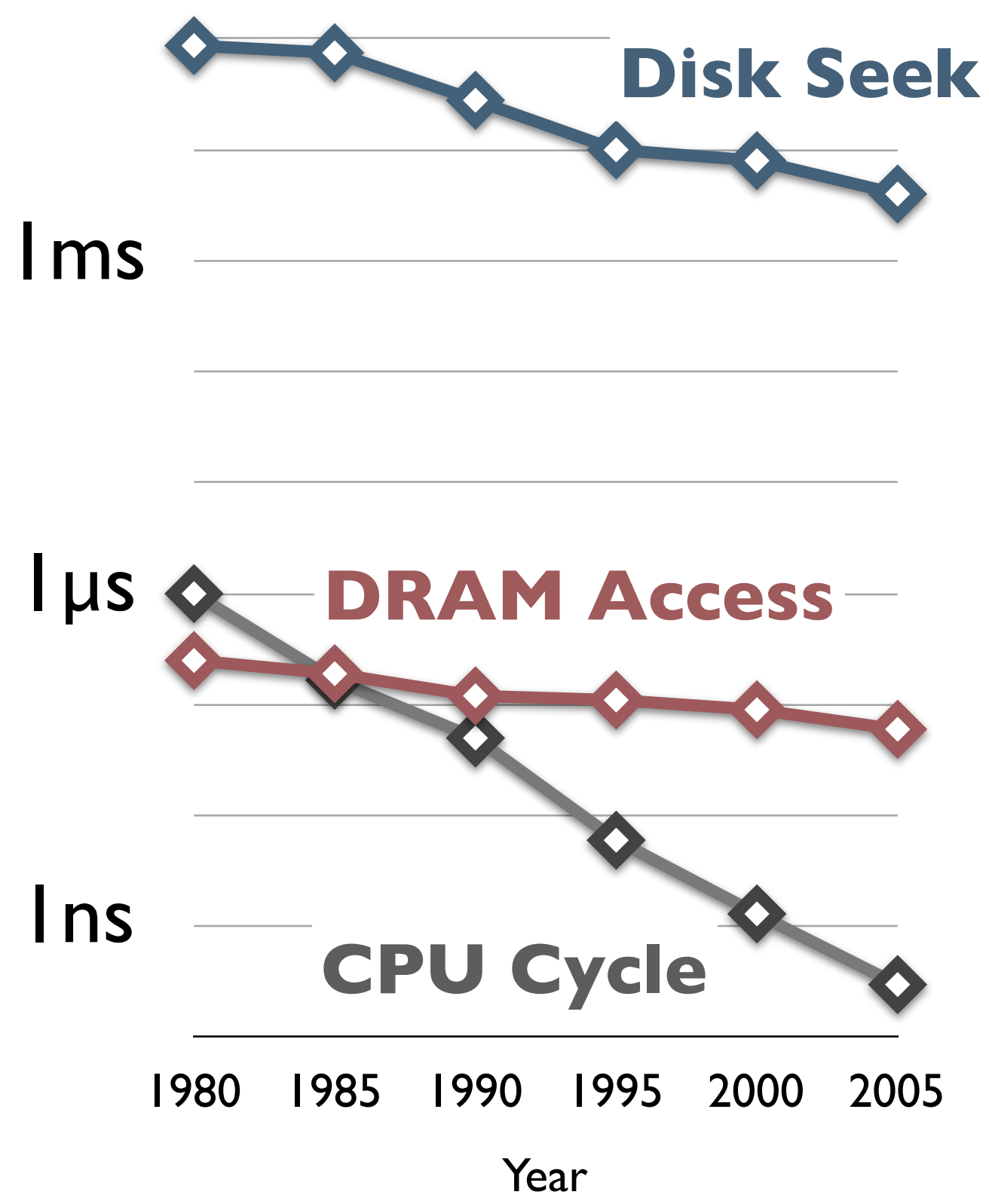


Gigahertz is not free

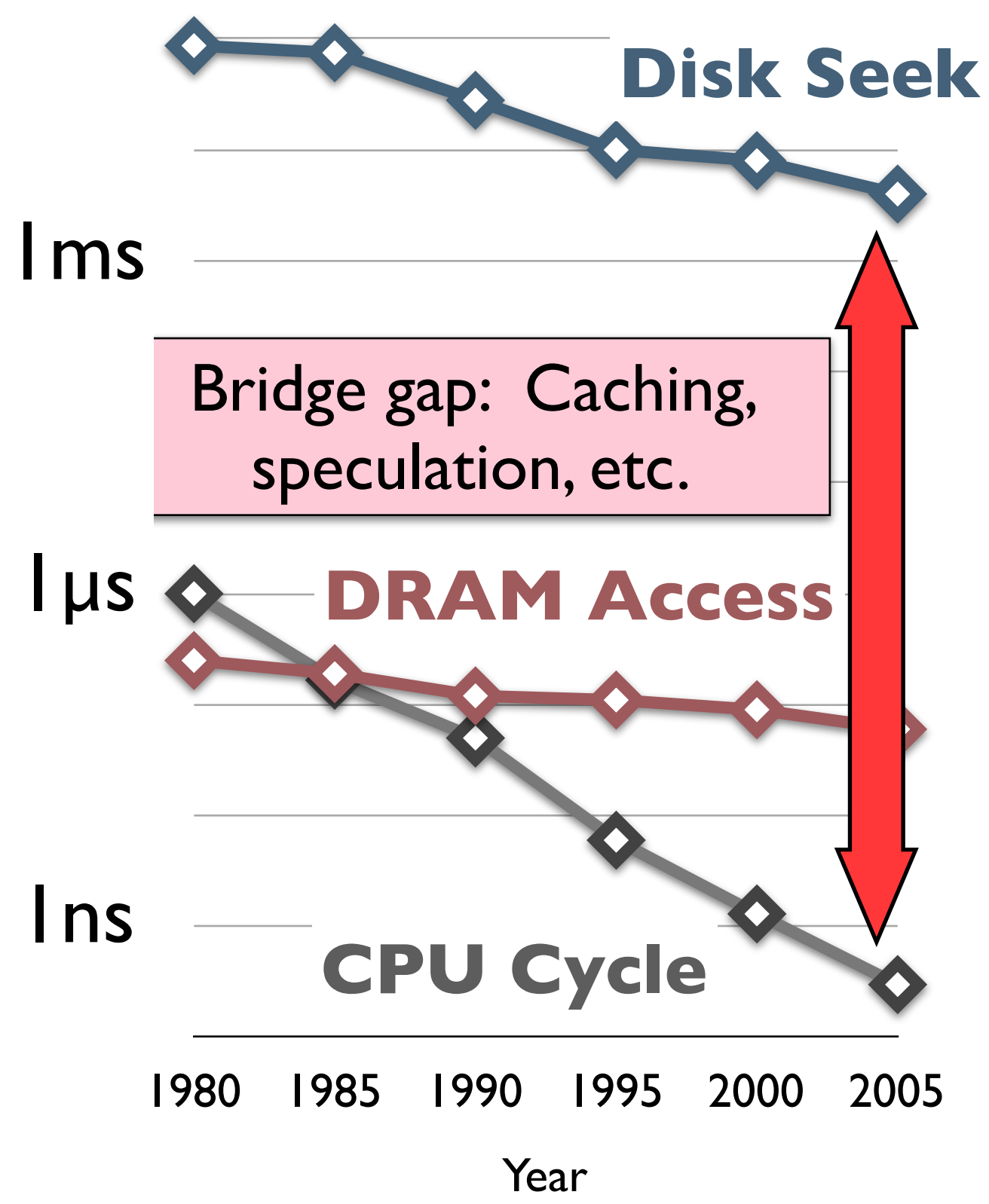
Speed and power calculated from specification sheets
Power includes “system overhead” (e.g., Ethernet)



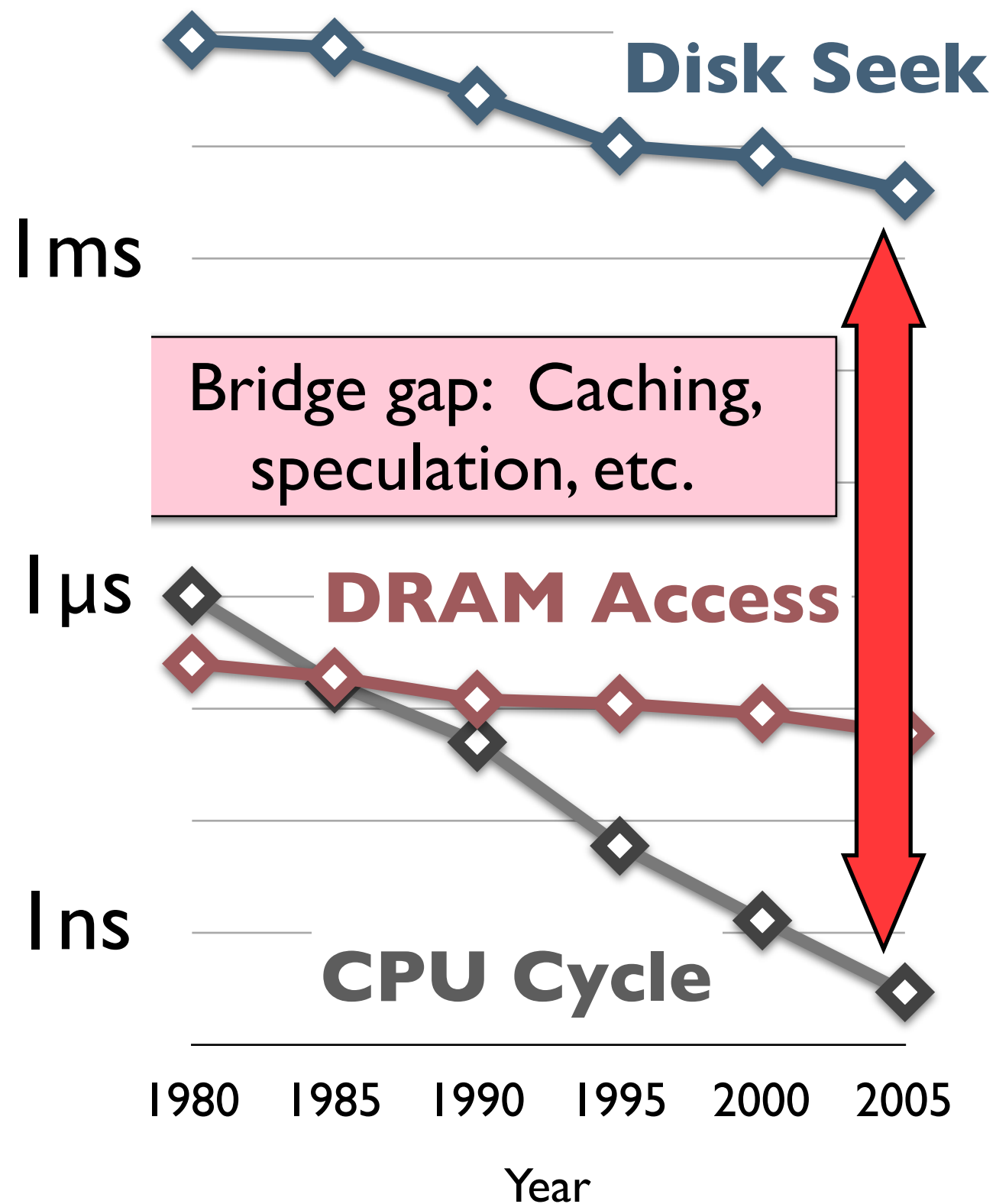
The Memory Wall



The Memory Wall

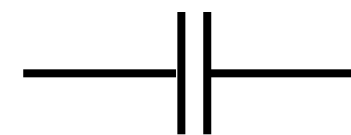


The Memory Wall

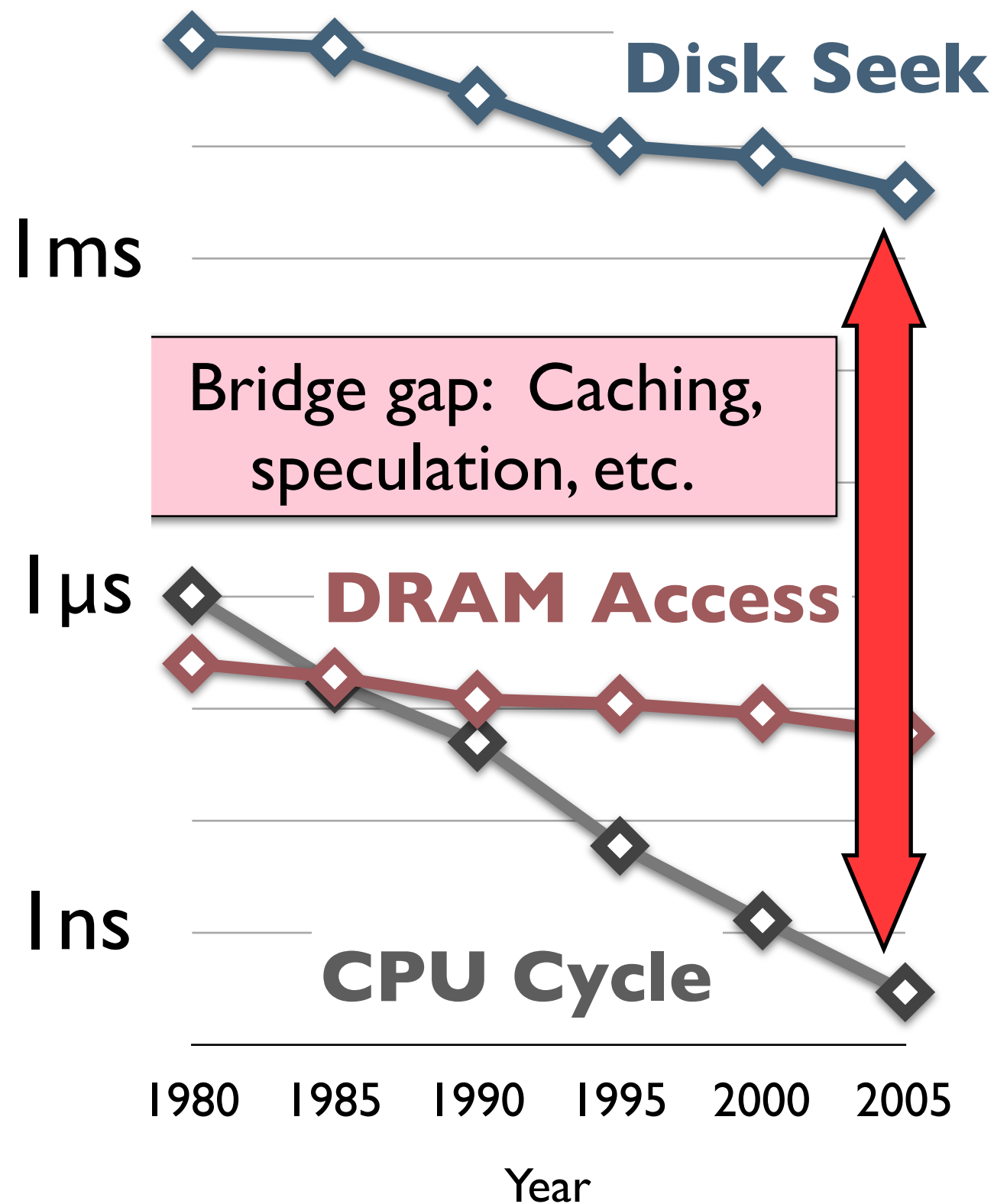


Transistors

Have the soul of a capacitor

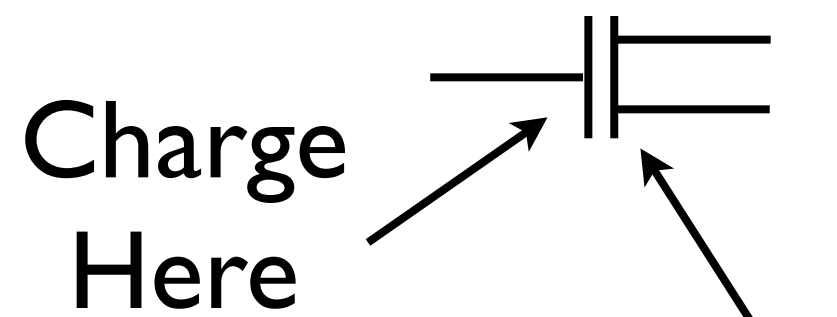


The Memory Wall



Transistors

Have the soul of a capacitor



Moves charge carriers here

Which lets current flow

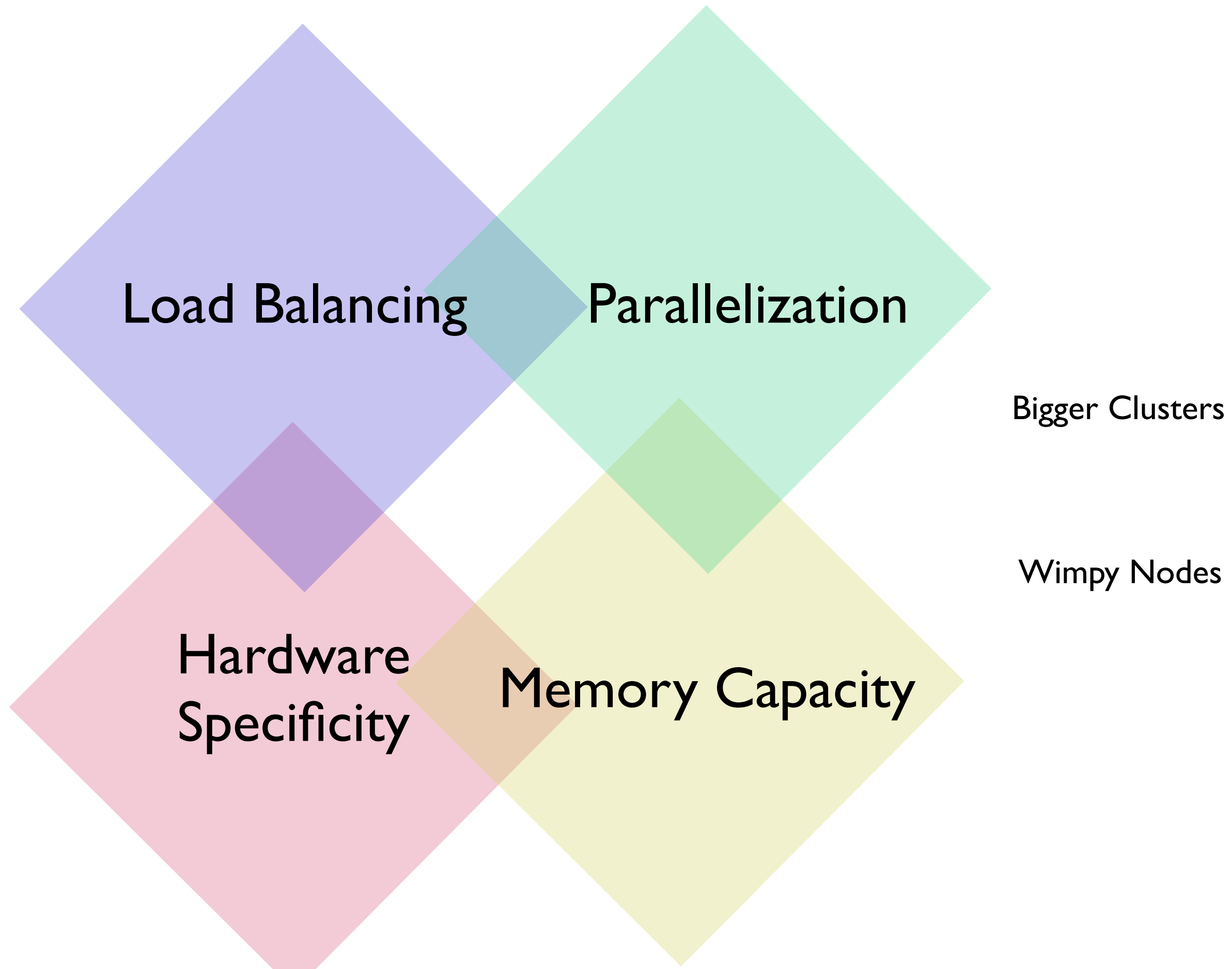
Gigahertz hurts

Remember:
Memory capacity costs you

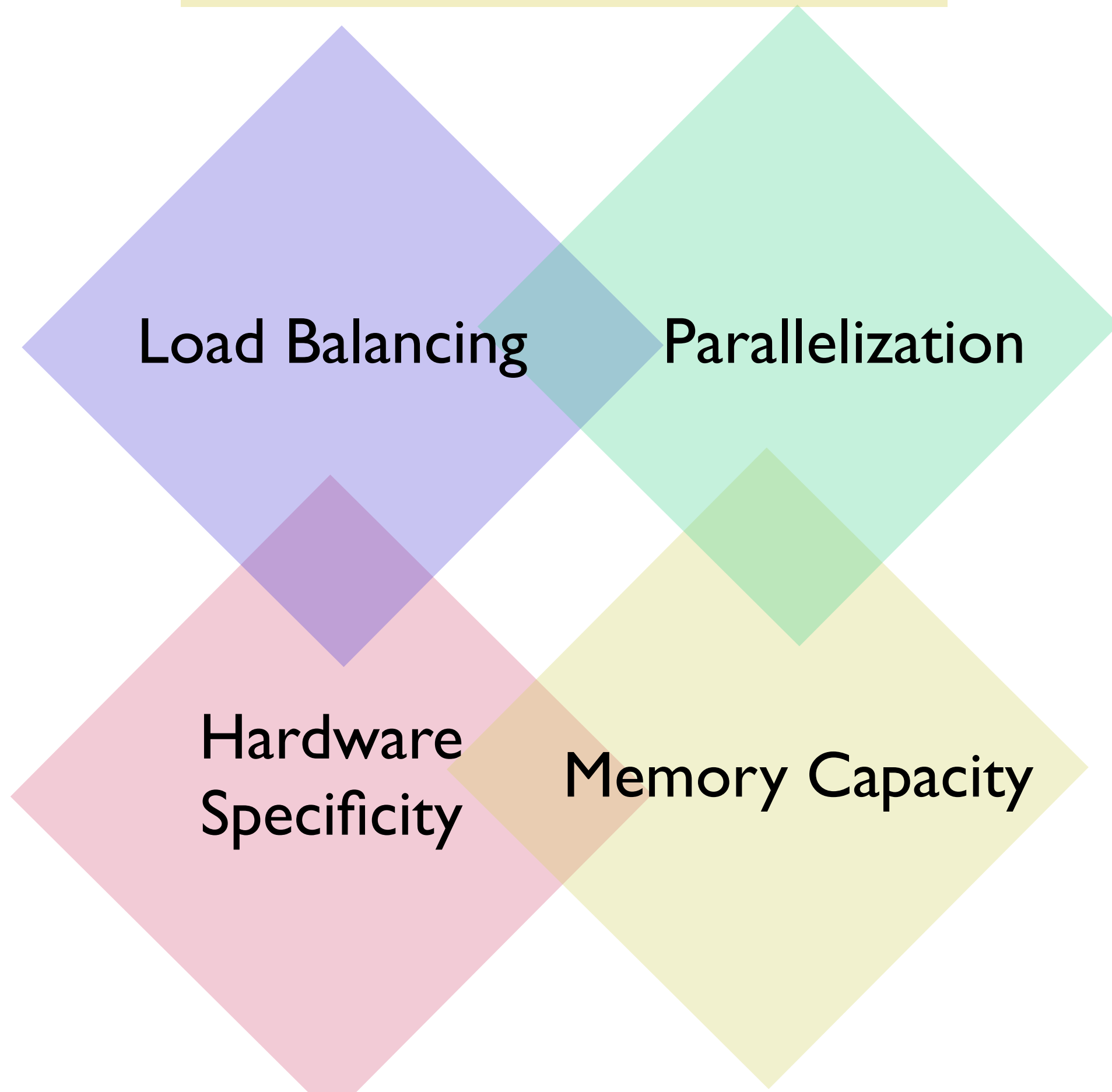
“Wimpy” Nodes

1.6 GHz Dual-core Atom
32-160 GB Flash SSD
Only 1 GB DRAM!

“Each decimal order of magnitude increase in parallelism requires a major redesign and rewrite of parallel code” - Kathy Yelick



The FAWN Quad of Pain

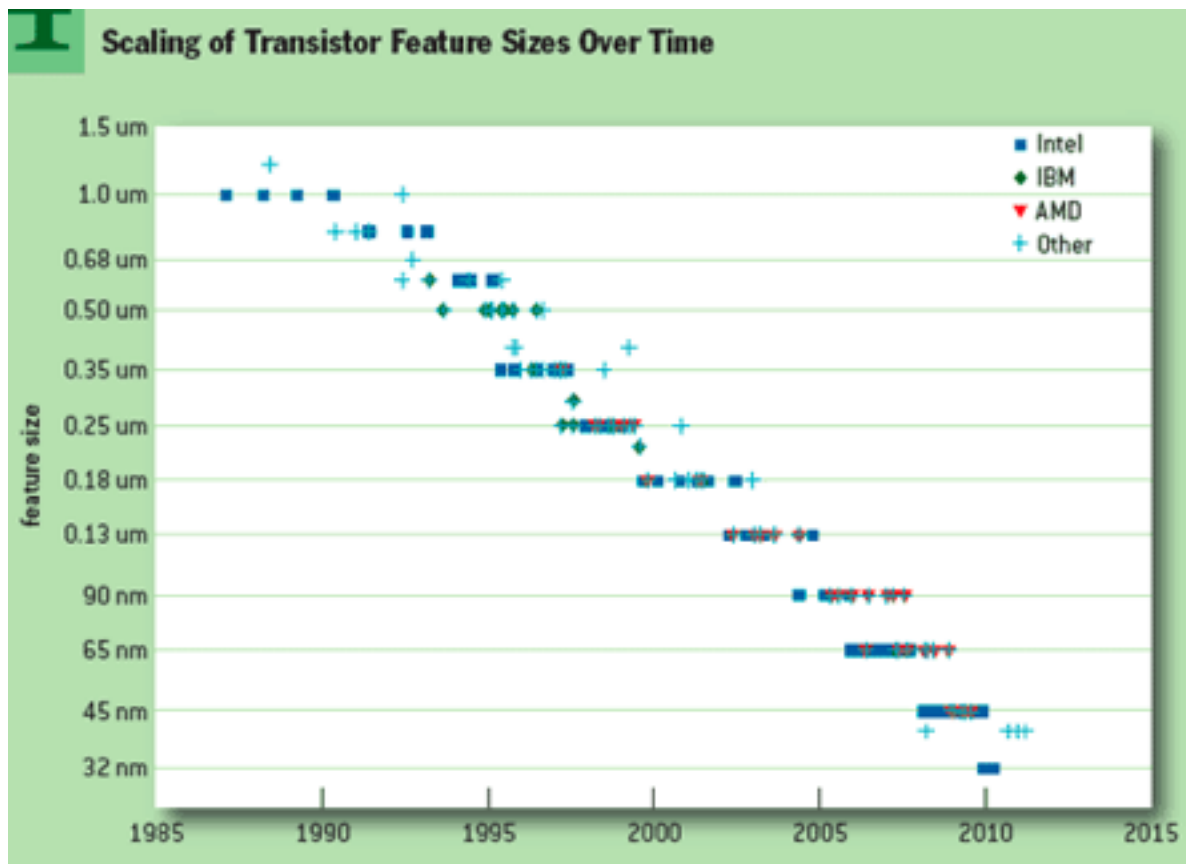


Bigger Clusters

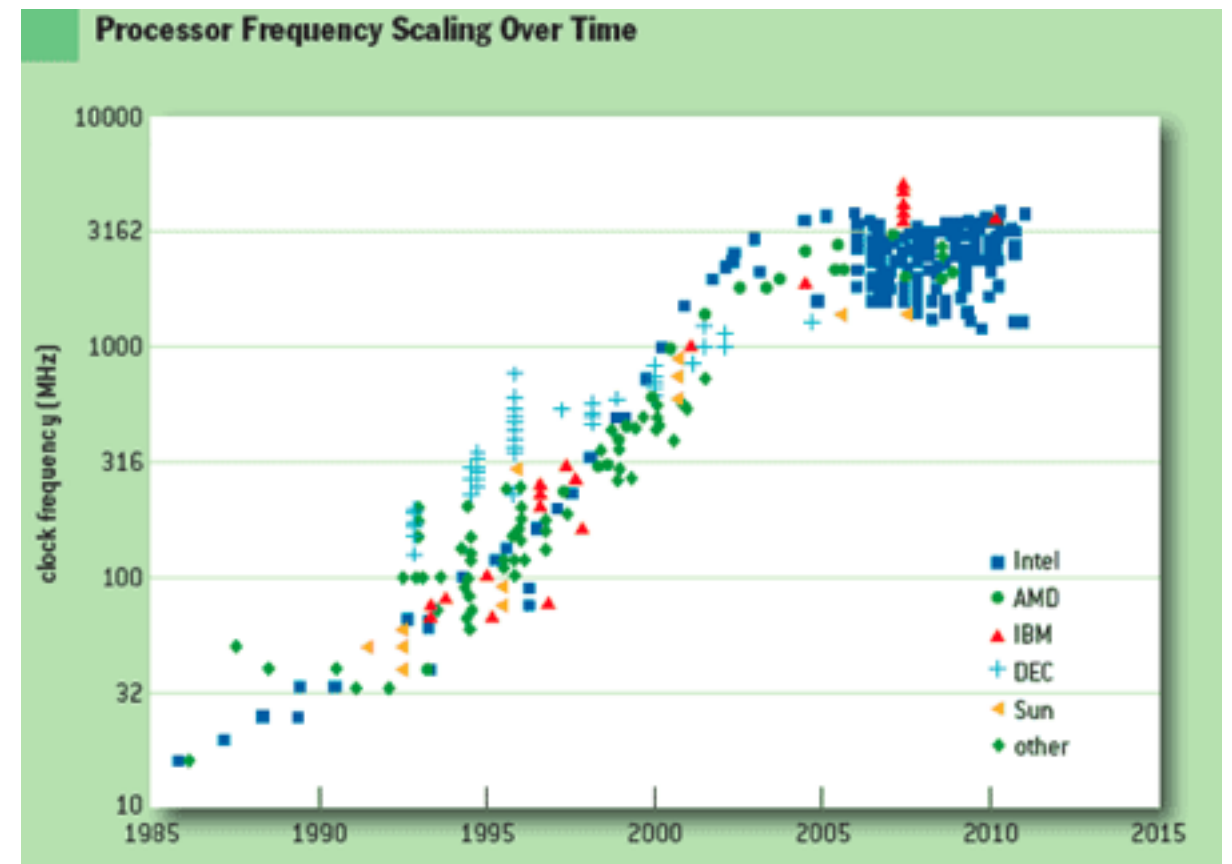
Wimpy Nodes

It's not just masochism

Moore



Dennard



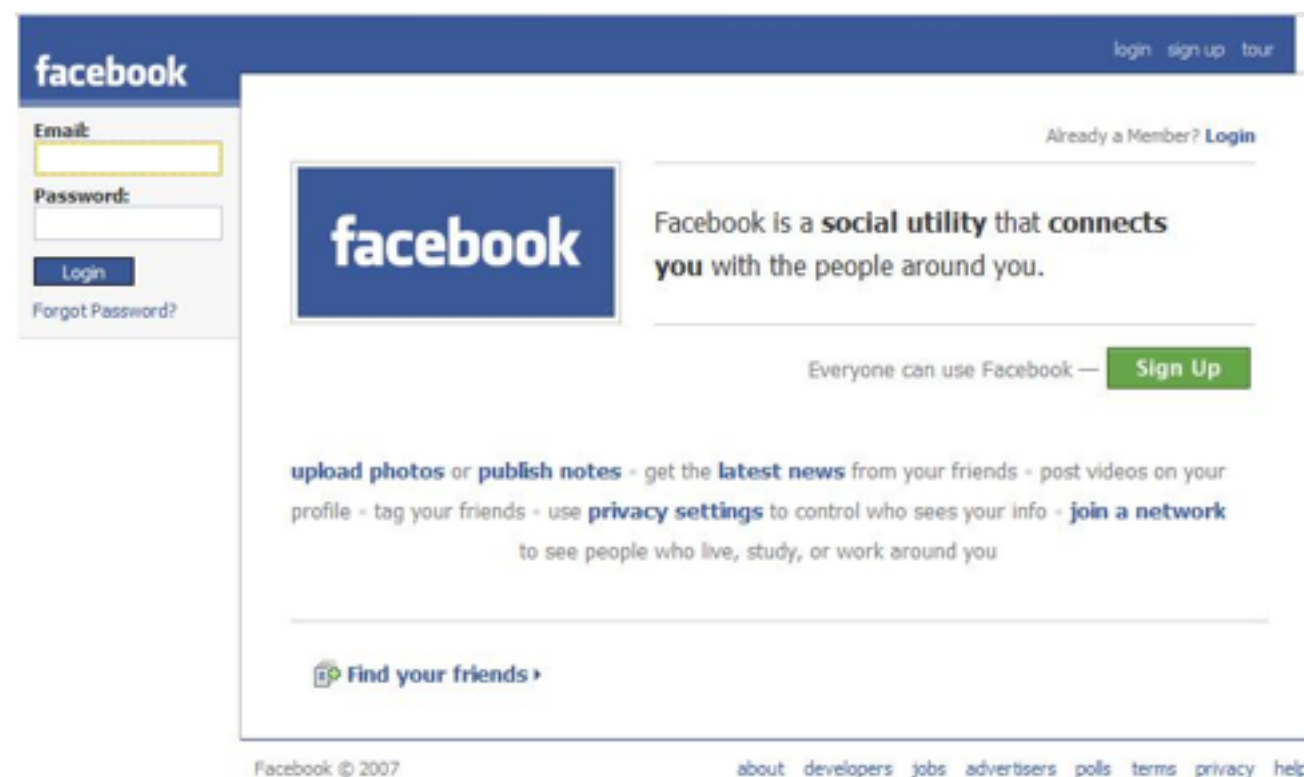
(Figures from Danowitz, Kelley, Mao, Stevenson, and Horowitz: CPU DB)

All systems will face this challenge over time

FAWN:
It started
with a key-value store

Key-value storage systems


- Critical infrastructure service
- Performance-conscious
- Random-access, read-mostly, hard to cache




Small record, random access

99 friends


See All




Carsten Varming




Timor Tsentsiper




Arvind Chari



Corey Iyican




John Bethencourt



Ram Ravichandran

Create a Profile Badge


Sep 21



Dan Wendlandt wrote at 6:47pm

have a good one man. hope the facebook TG was fun, the email was hilarious

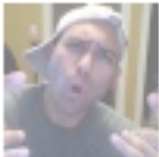
Wall-to-Wall – Write on Dan's Wall



Patrick Gage Kelley wrote at 2:42pm

Oh! birthday!


Wall-to-Wall – Write on Patrick's Wall



Jagan Seshadri wrote at 1:50pm

Happy birthday Vij! 24 and there's so much more...

Wall-to-Wall – Write on Jagan's Wall




Vish Subramanian wrote at 3:48am

happy birthday dude, its been awhile!

Wall-to-Wall – Write on Vish's Wall

Sep 19



Bobby Gregg wrote at 2:22pm

hi vijay! i'm super early but i'm bad about checking facebook regularly nowadays so i wanted to say happy birthday. let's catch up about our respective grad school woes.

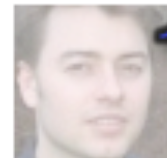
Wall-to-Wall – Write on Bobby's Wall

13

Small record, random access

```
Select name,photo from users where uid=513542;
```

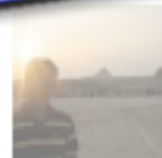
99 friends



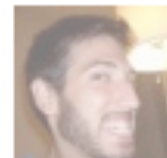
Carsten
Varming



Timor
Tsentsiper



Arvind
Chari



Corey
Iyican



John
Bethencourt



Ram
Ravichandran

Create a Profile Badge



Dan Wendlandt wrote at 6:47pm

have a good one man. hope the facebook TG was fun, the email was hilarious

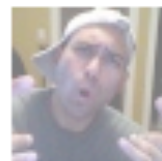
Wall-to-Wall – Write on Dan's Wall



Patrick Gage Kelley wrote at 2:42pm

Oh! birthday!

Wall-to-Wall – Write on Patrick's Wall



Jagan Seshadri wrote at 1:50pm

Happy birthday Vij! 24 and there's so much more...

Wall-to-Wall – Write on Jagan's Wall



Vish Subramanian wrote at 3:48am

happy birthday dude, its been awhile!

Wall-to-Wall – Write on Vish's Wall

Sep 19




Bobby Gregg wrote at 2:22pm

hi vijay! i'm super early but i'm bad about checking facebook regularly nowadays so i wanted to say happy birthday. let's catch up about our respective grad school woes.


Wall-to-Wall – Write on Bobby's Wall

Small record, random access

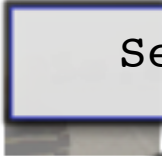
99 friends See All Sep 21




Carsten Varming




Timor Tsentsiper




Arvind Chari




Corey Iyican



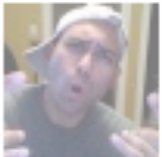
John Bethencourt




Ram Ravichandran



Patrick Gage Kelley wrote at 2:42pm
Oh! birthday!
Wall-to-Wall – Write on Patrick's Wall




Jagan Seshadri wrote at 1:50pm
Happy birthday Vij! 24 and there's so much more...
Wall-to-Wall – Write on Jagan's Wall



Vish Subramanian wrote at 3:48am
happy birthday dude, its been awhile!
Wall-to-Wall – Write on Vish's Wall

Sep 19



Bobby Gregg wrote at 2:22pm
hi vijay! i'm super early but i'm bad about checking facebook regularly nowadays so i wanted to say happy birthday. let's catch up about our respective grad school woes.
Wall-to-Wall – Write on Bobby's Wall

Create a Profile Badge

Select name, photo from users where uid=818503;

Small record, random access

99 friends See All

Carsten Varming Timor Tsentsiper Arvind Chari

Corey Iyca John Bethenco Ram Ravichandran

Sep 21

Dan Wendlandt wrote at 6:47pm
have a good one man. hope the facebook TG was fun, the email was hilarious
Wall-to-Wall – Write on Dan's Wall

Patrick Gage Kelley wrote at 2:42pm
Oh! birthday!
Wall-to-Wall – Write on Patrick's Wall

Jagan Seshadri wrote at 1:50pm
e's so much more...
Wall

Create a Profile Badge

Sep 19

Vish Subramanian wrote at 3:48am
happy birthday dude, its been awhile!
Wall-to-Wall – Write on Vish's Wall

Bobby Gregg wrote at 2:22pm
hi vijay! i'm super early but i'm bad about checking facebook regularly nowadays so i wanted to say happy birthday. let's catch up about our respective grad school woes.
Wall-to-Wall – Write on Bobby's Wall

Select name,photo from users where uid=468883;

Small record, random access

The image shows a screenshot of a Facebook profile page. On the left, there is a section titled "99 friends" with a "See All" link. Below this, there are six small profile pictures of friends, each with their name underneath: Carsten Varming, Timor Tsentsiper, Arvind Chari, Corey Iyican, John Bethencourt, and Rajendra. On the right, there is a section titled "Sep 21" showing a list of posts. The first post is by Dan Wendlandt, dated at 6:47pm, with the text "have a good one man. hope the facebook TG was fun, the email was hilarious". The second post is by Patrick Gage, dated at 2:42pm, with the text "Oh! birthday!". The third post is by Vish Subramanian, dated at 3:48am, with the text "happy birthday dude, its been awhile!". The fourth post is by Bobby Gregg, dated at 2:22pm, with the text "hi vijay! i'm super early but i'm bad about checking facebook regularly nowadays so i wanted to say happy birthday. let's catch up about our respective grad school woes.".

Select wallpost from posts where pid=13821828188;

Select name,photo from users where uid=124111;

Create a Profile Badge

Small record, random access

99 friends

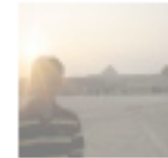
See All



Carsten



Arvind



Corey



Dan

have a good one man. hope the facebook TG was fun, the email was hilarious



Oh! birthday!

Wall-to-Wall - Write on Patrick's Wall

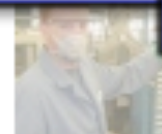


Wall-to-Wall - Write on Vish's Wall

Vish

happy birthday dude, its been awhile!

Wall-to-Wall - Write on Vish's Wall



hi vijay! i'm sup about checking facebook regularly nowadays so i wanted to say happy birthday. let's catch up about our respective grad school woes.

Wall - Write on Bobby's Wall

Select wallpost from posts where pid=89888333522;

Select wallpost from posts where pid=13821828188;

Select name,photo from users where uid=474488;

Select name,photo from users where uid=124566;

Select name,photo from users where uid=124111;

Select name,photo from users where uid=12223;

Select wallpost from posts where pid=12314144887;

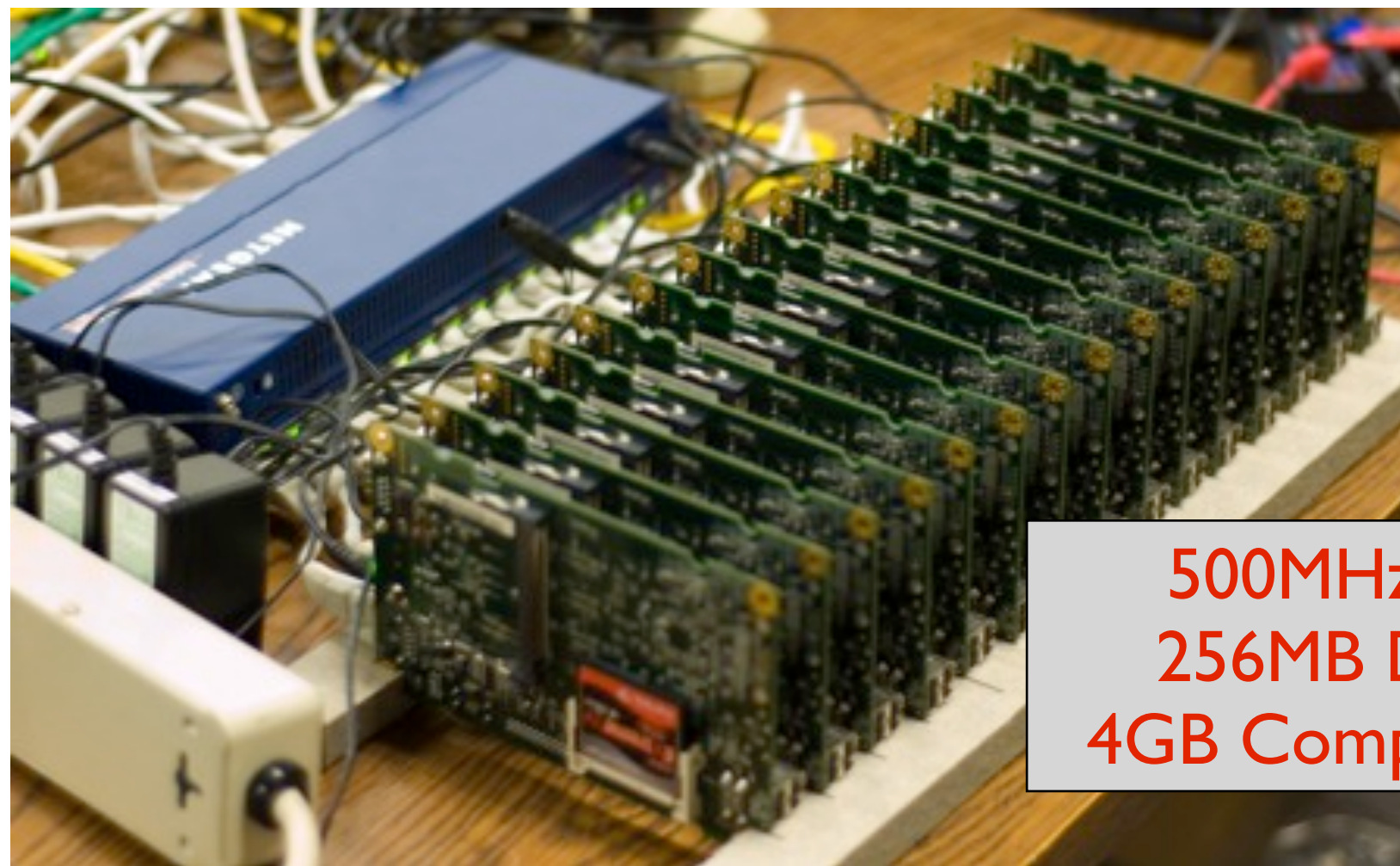
Select name,photo from users where uid=997788;

Select wallpost from posts where pid=738838402;

Select name,photo from users where uid=357845;

FAWN-DS and -KV: Key-value Storage System

Goal: improve **Queries/Joule**



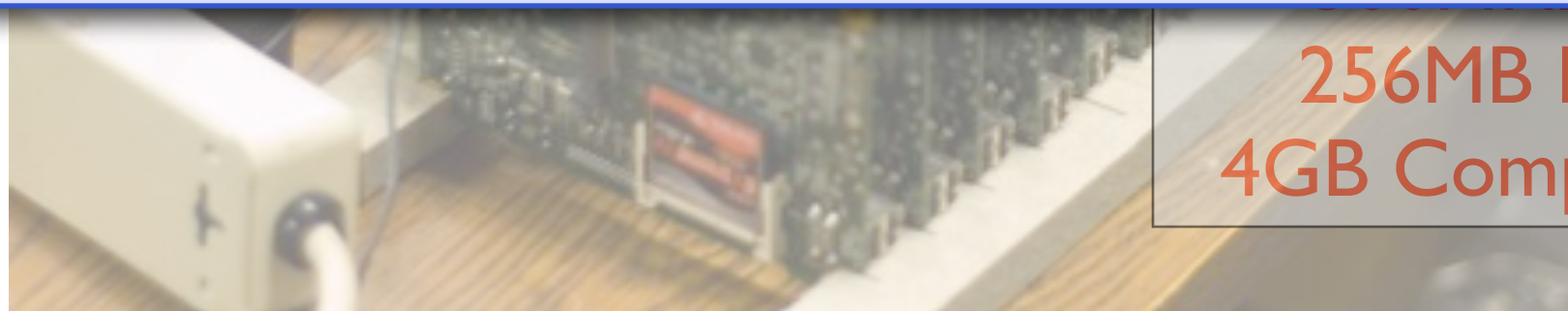
500MHz CPU
256MB DRAM
4GB CompactFlash

FAWN-DS and -KV: Key-value Storage System

Goal: improve **Queries/Joule**

Unique Challenges:

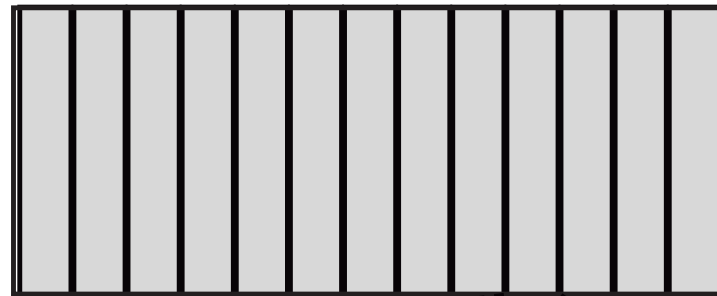
- Wimpy CPUs, limited DRAM
- Flash poor at small random writes
- Sustain performance during membership changes



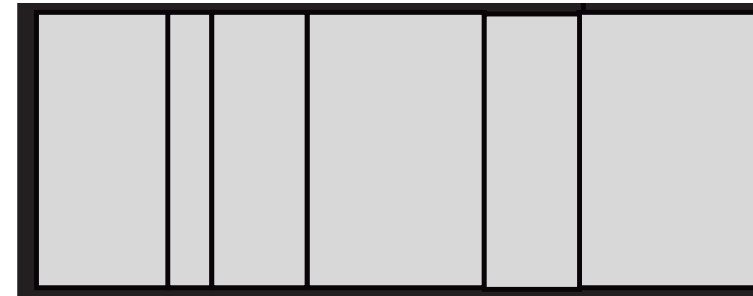
256MB DRAM
4GB CompactFlash

Avoiding random writes

Hashtable



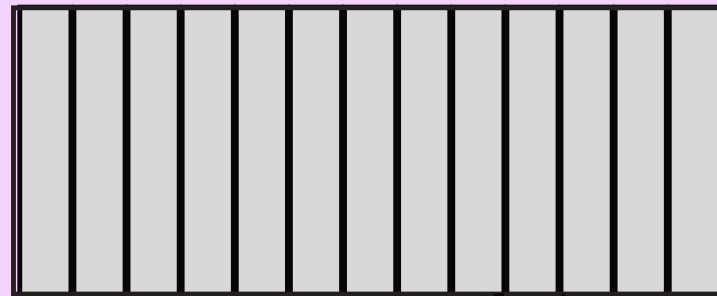
Data region



Avoiding random writes

In DRAM

Hashtable



In Flash

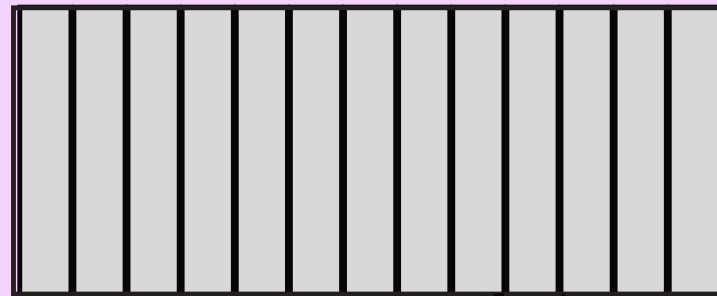
Data region



Avoiding random writes

In DRAM

Hashtable



In Flash

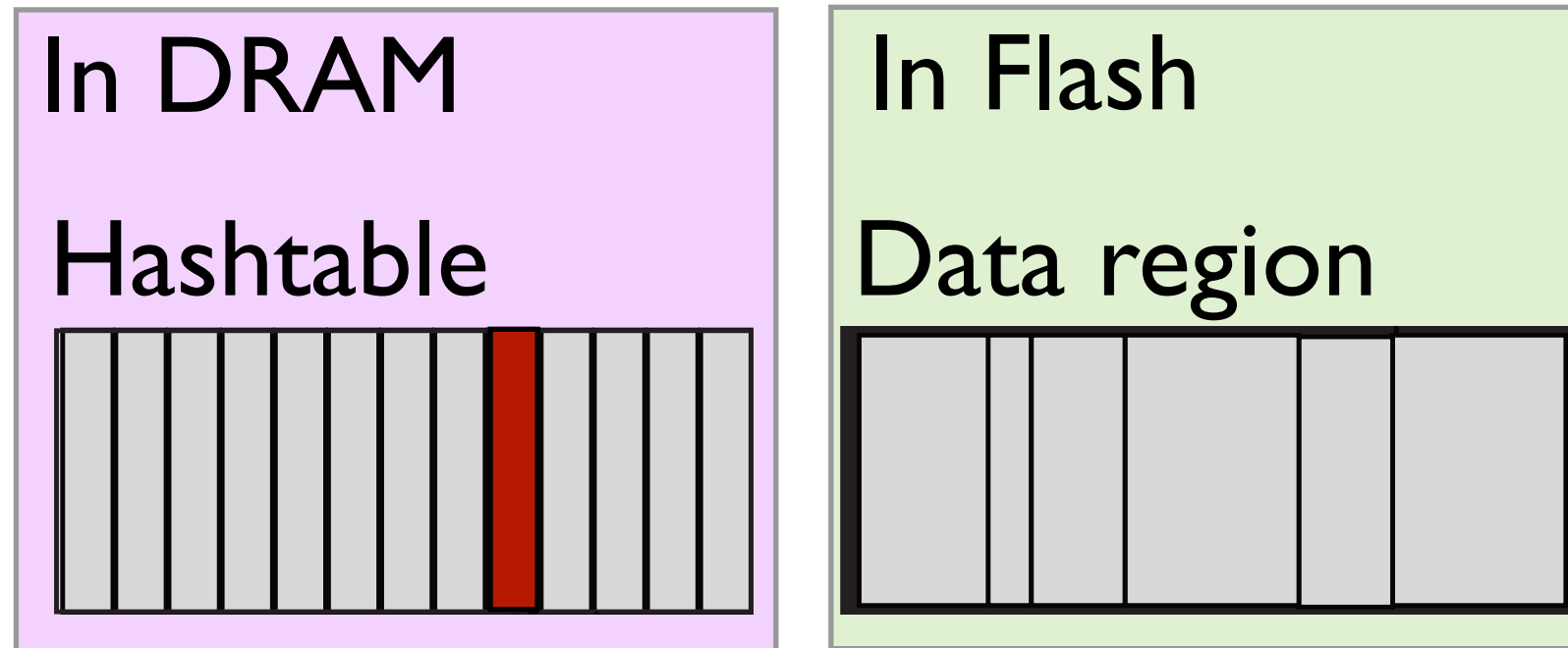
Data region



Put

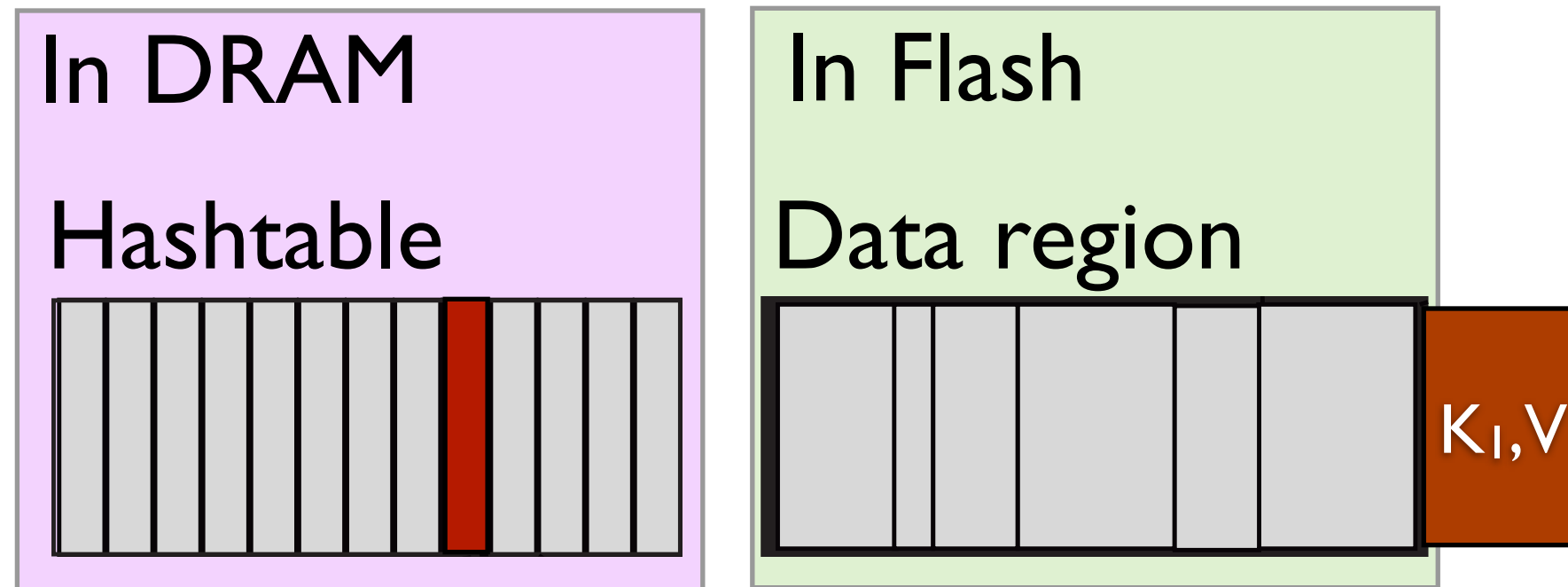


Avoiding random writes



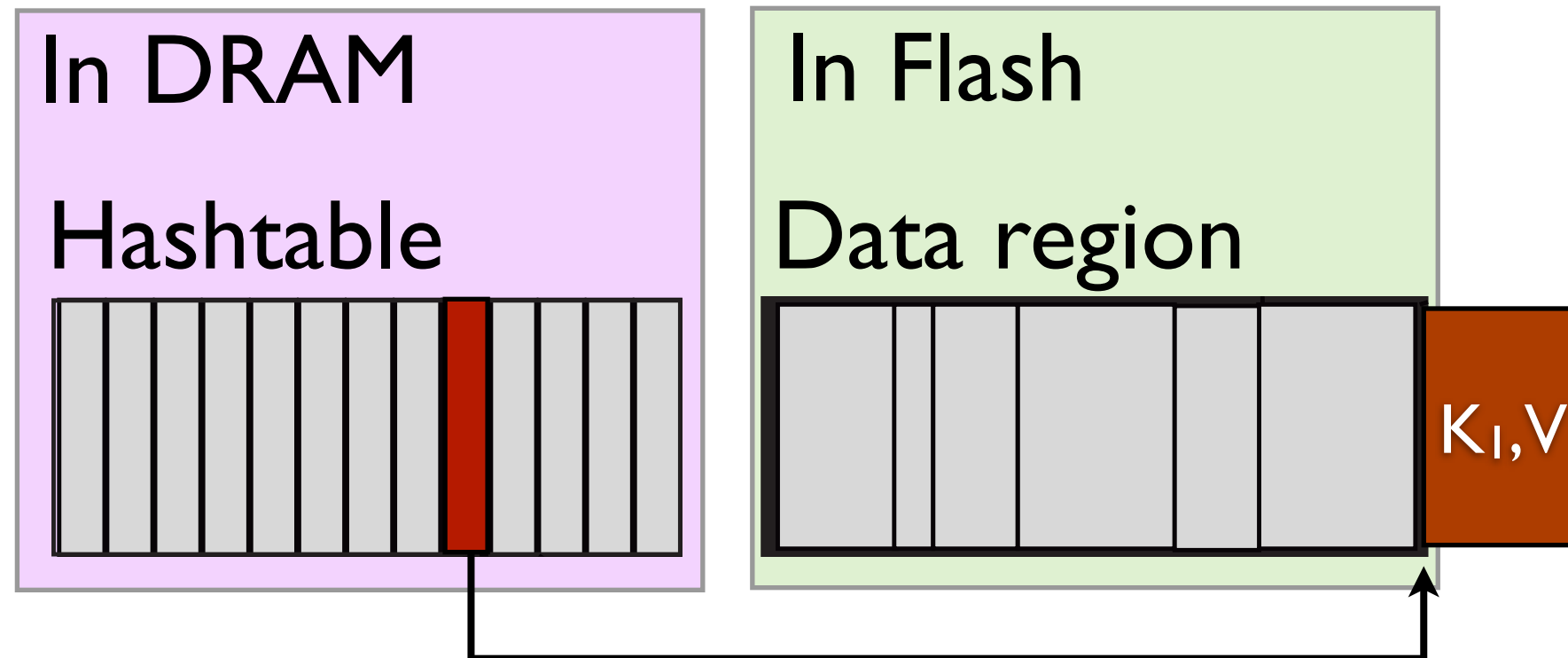
Put K_i, V

Avoiding random writes



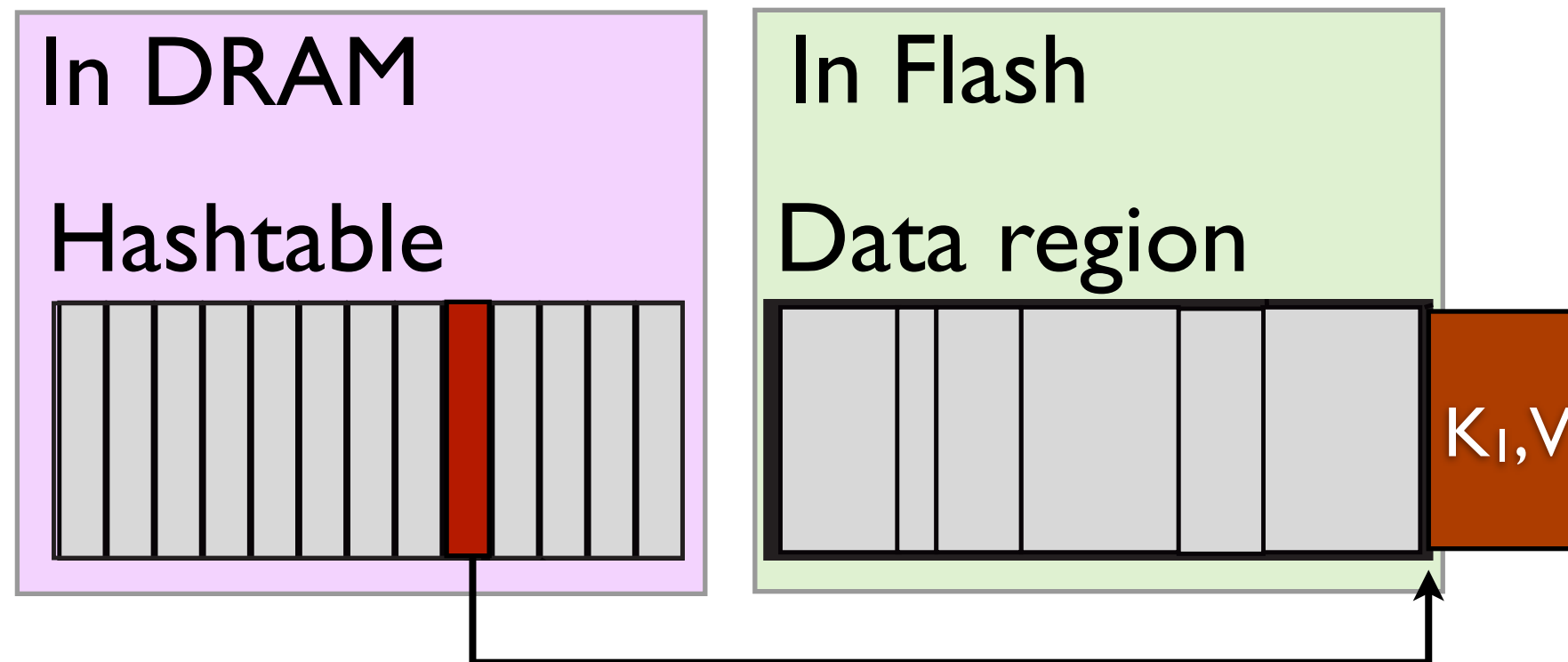
Put

Avoiding random writes



Put

Avoiding random writes



Put

All writes to Flash are sequential

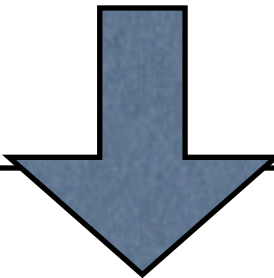
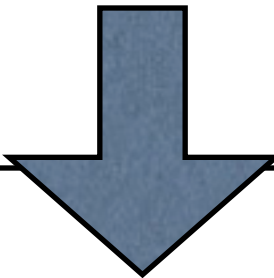
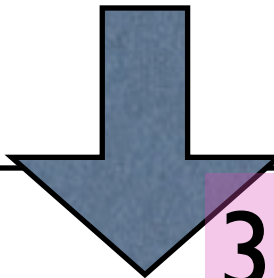
Research Example

- Developed DRAM-efficient system to find location on flash
 - (“Partial-key hashing”) 2008-9
- We’ve continued this since then:
 - Partial-key cuckoo hashing 2011
 - Optimistic concurrent cuckoo hashing 2012

Evaluation Takeaways

- 2008: FAWN-based system 6x more efficient than traditional systems
- Partial-key hashing enabled memory-efficient DRAM index for flash-resident data
- Can create high-performance, predictable storage service for small key-value pairs

And then we moved to Atom + SSD

| | | |
|---|---|---|
| Geode 500Mhz | 256MB | 4GB CF Card ~2k IOPS |
|  |  |  |
| Atom 1.6 Ghz single-core | 2GB | 120GB SSD ~60k IOPS |
| 6x | 8x | 30-60x |

FAWN-DS FAWN-KV

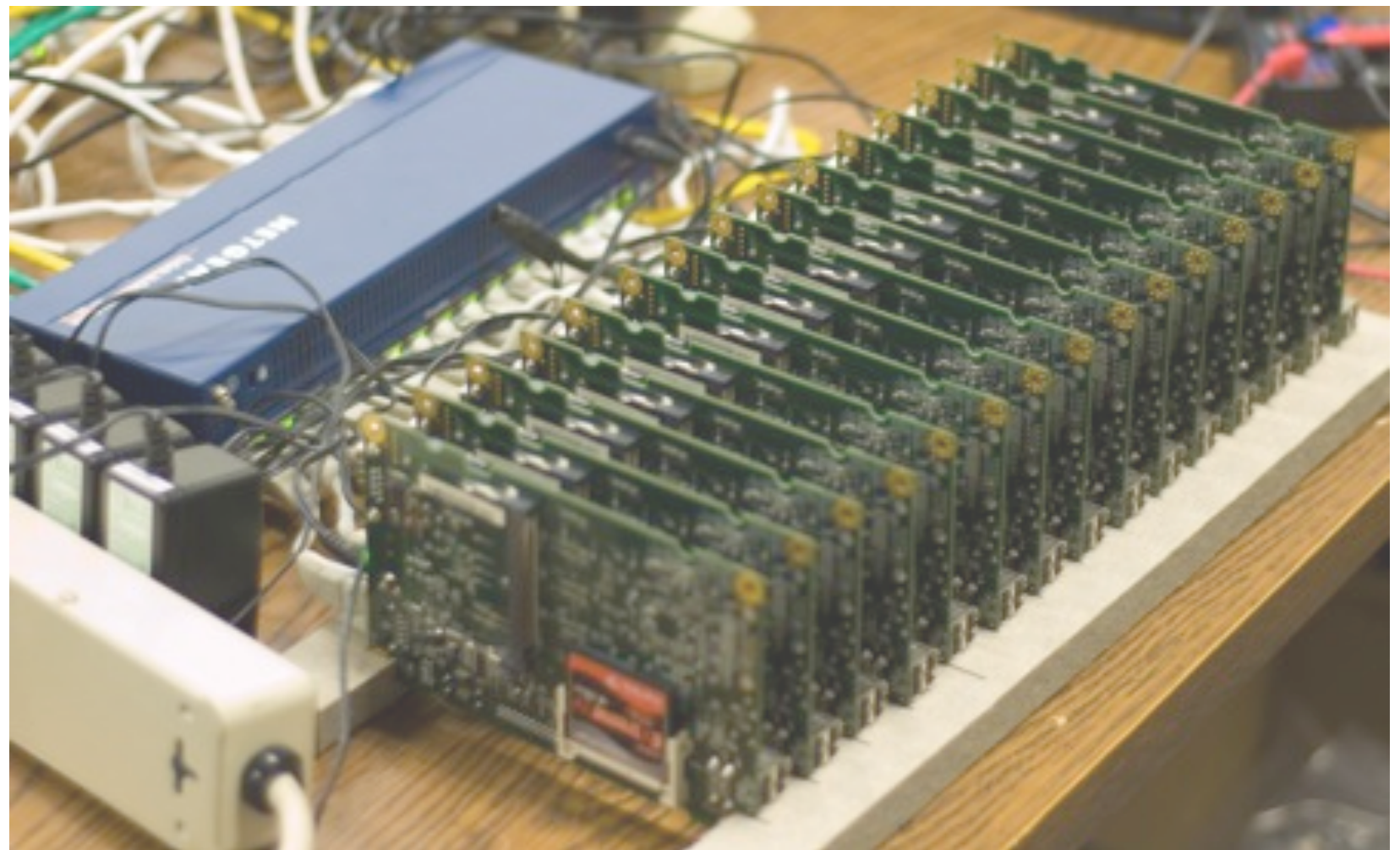
Small Cache Cuckoo

Fawn-KV

Fawn-DS

Fawn-DS

Fawn-DS



FAWN-DS FAWN-KV SILT Small Cache Cuckoo

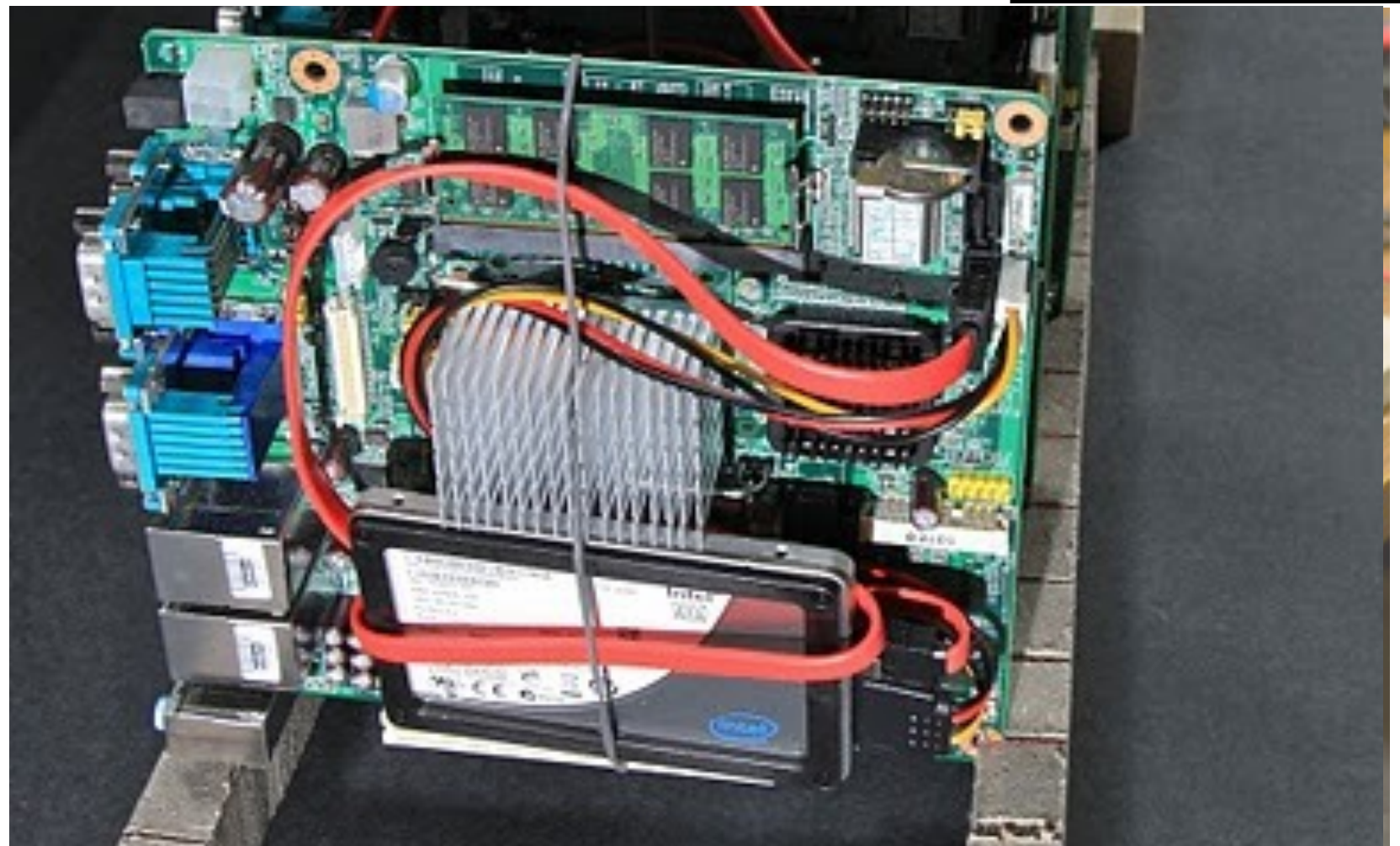
backend store
hyper-optimized
for low DRAM
and large flash

Fawn-KV

SILT

SILT

SILT



Systems begat
algorithms:

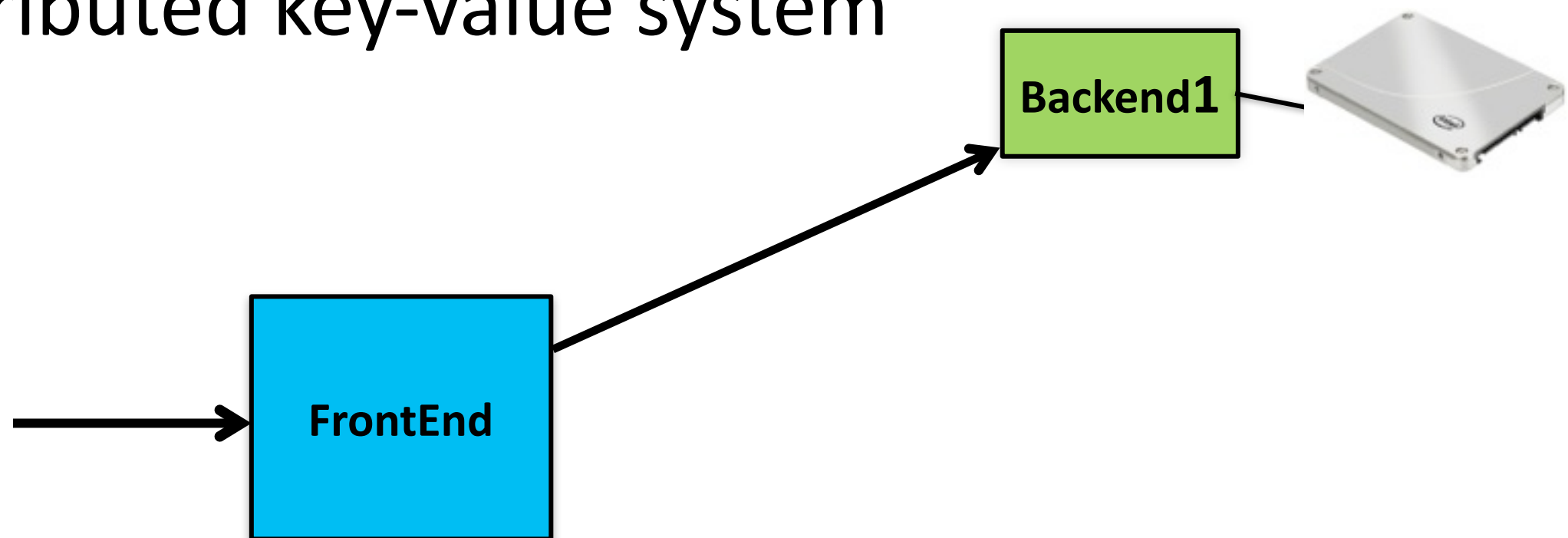
“Practical Batch-Updtable
External Hashing with Sorting”

H. Lim et al., **ALENEX** 2012

(Recently heard that Bing uses
several state-of-the-art,
memory-efficient indexes)

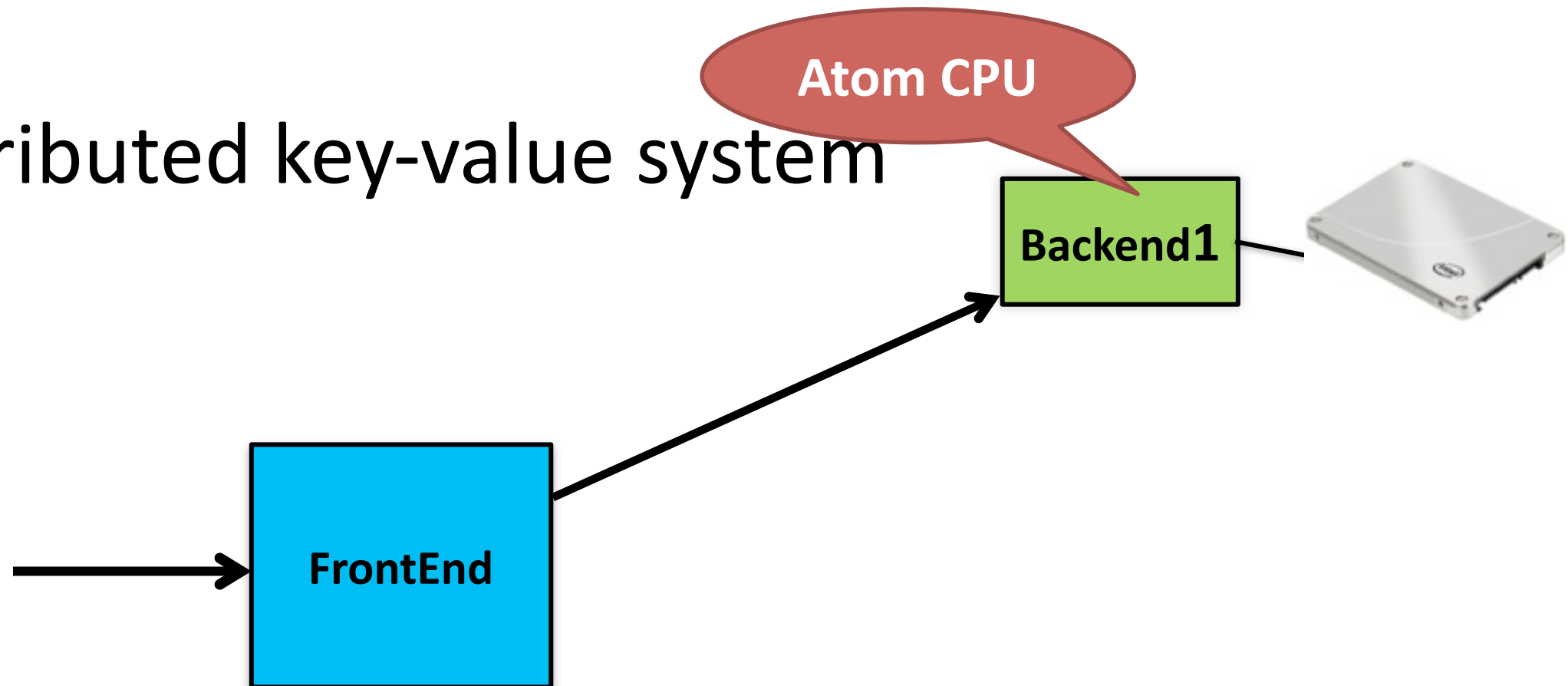
And now... Load imbalance

- Distributed key-value system



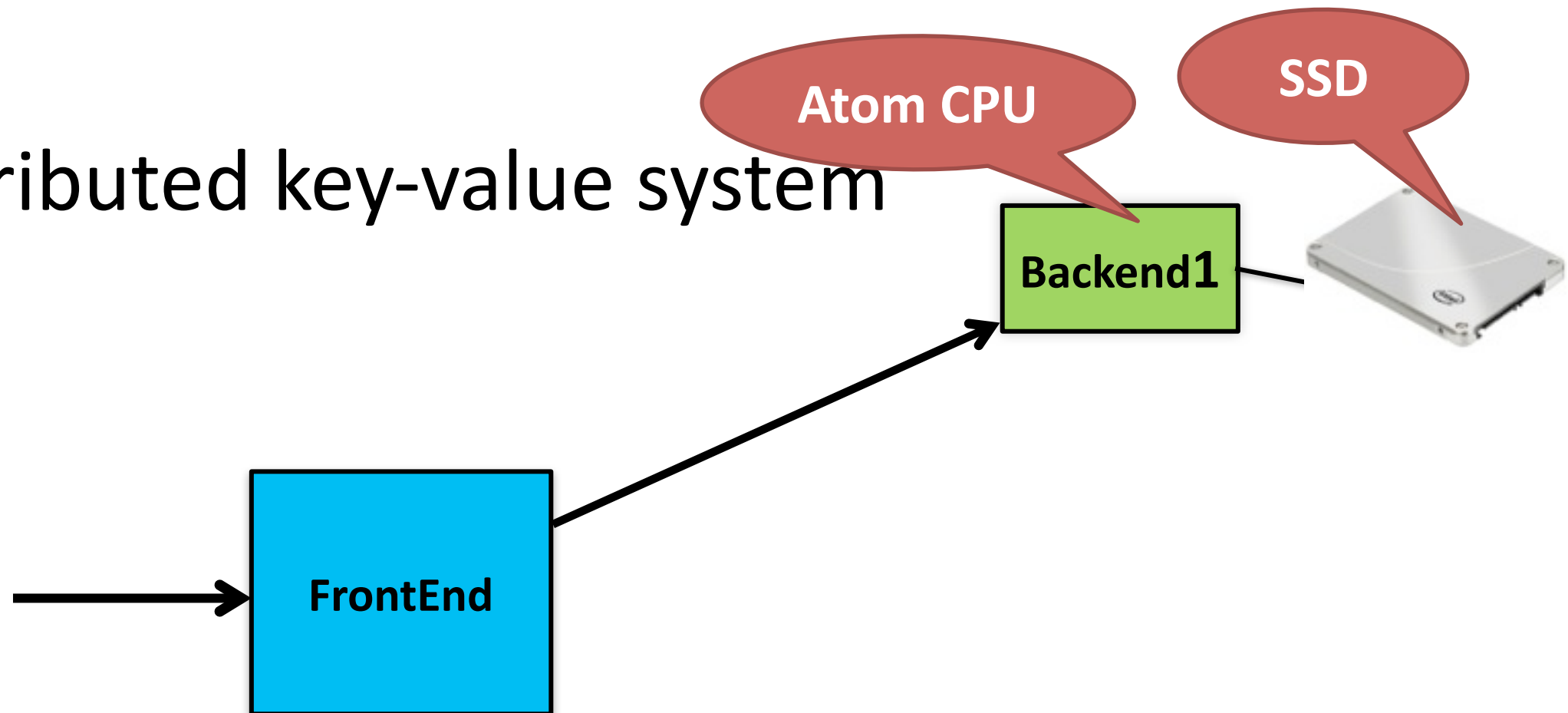
And now... Load imbalance

- Distributed key-value system



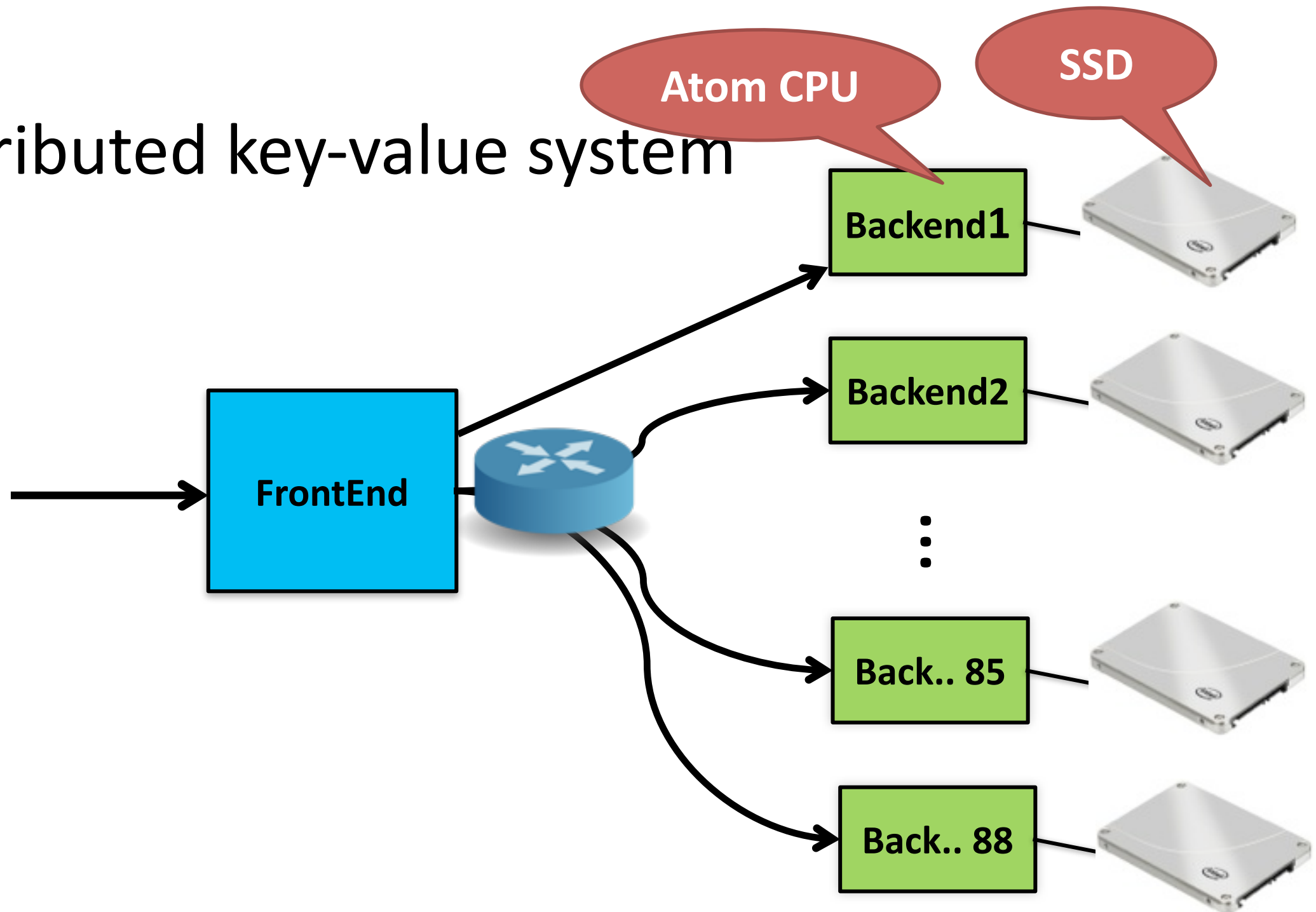
And now... Load imbalance

- Distributed key-value system



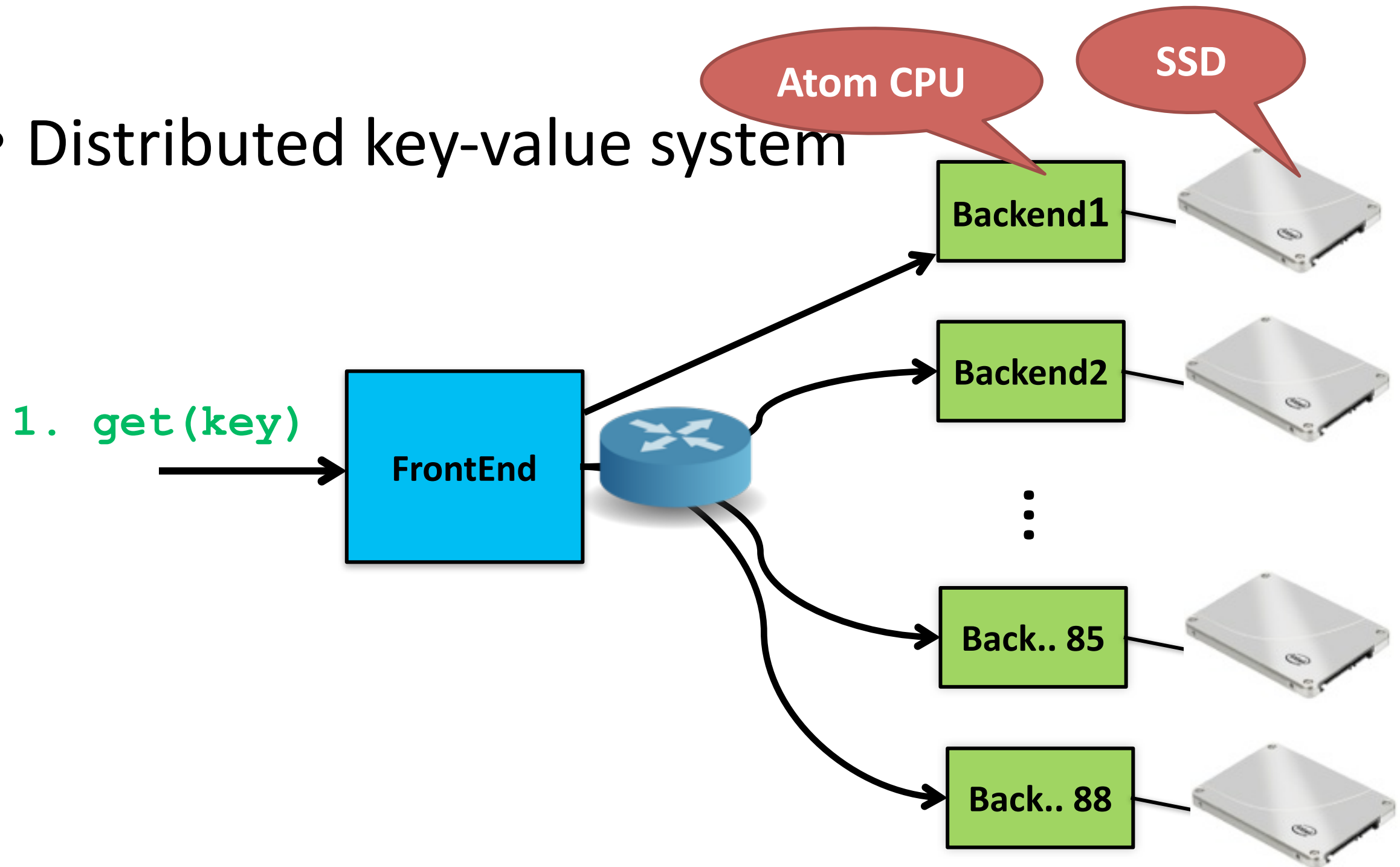
And now... Load imbalance

- Distributed key-value system



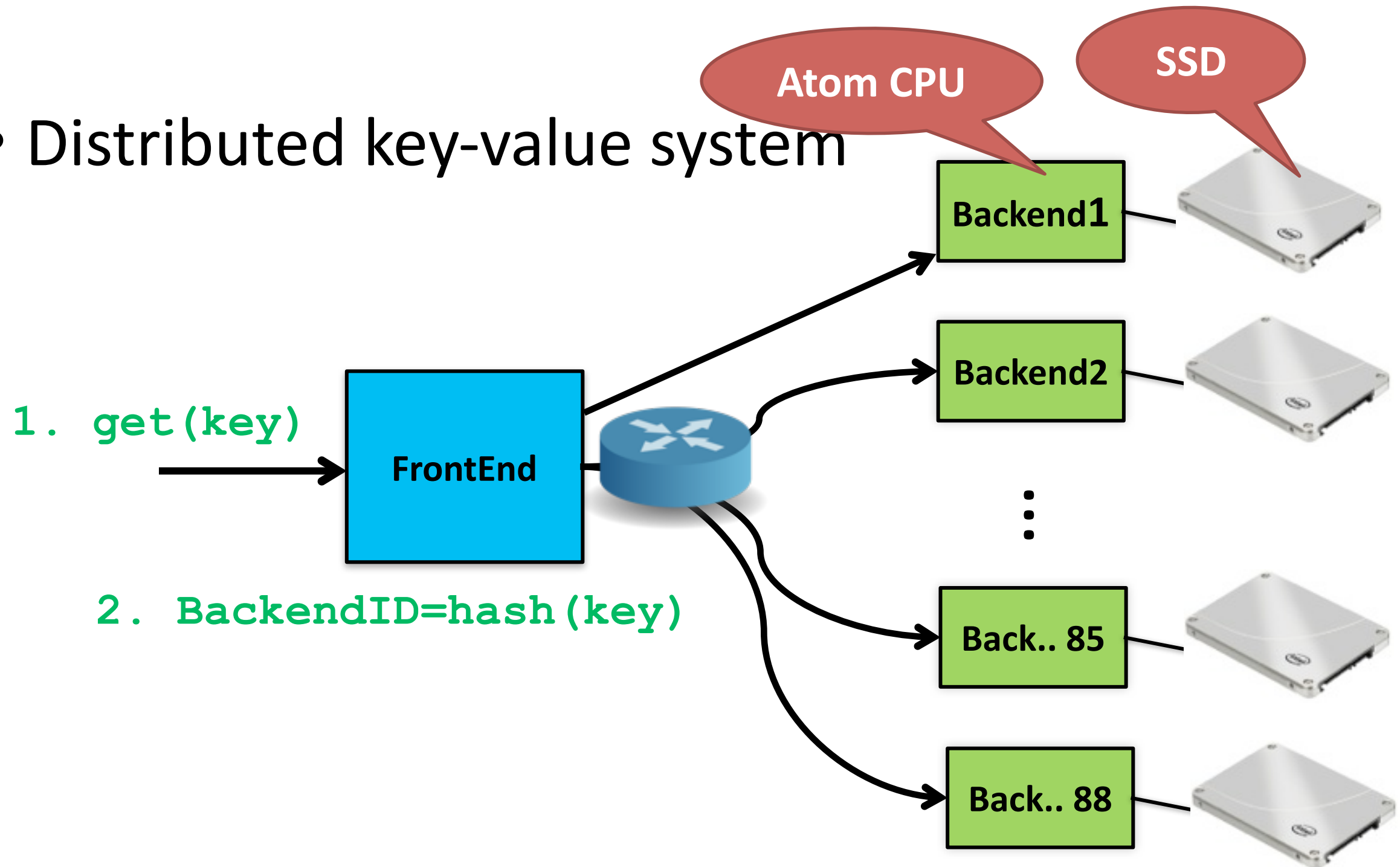
And now... Load imbalance

- Distributed key-value system



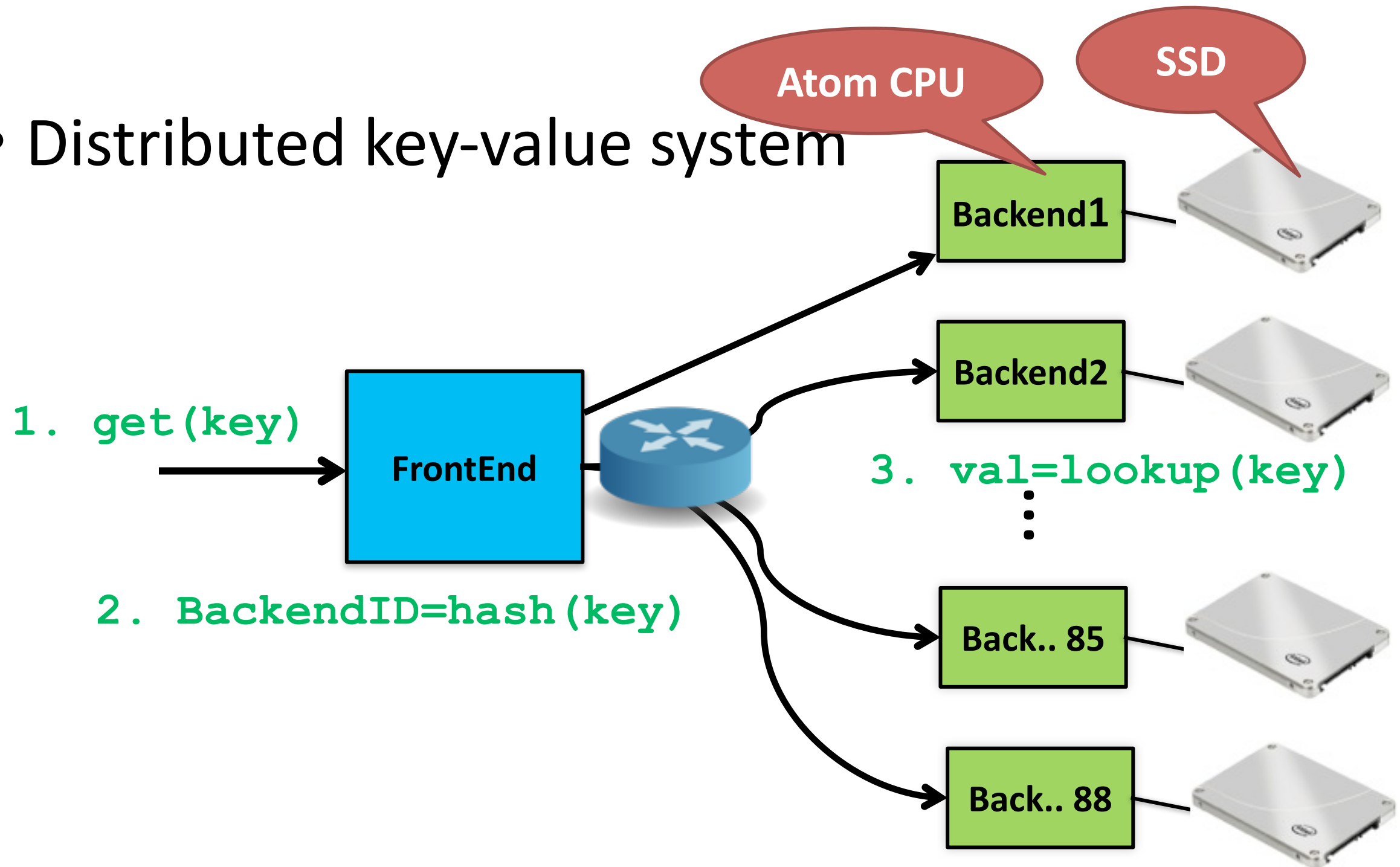
And now... Load imbalance

- Distributed key-value system



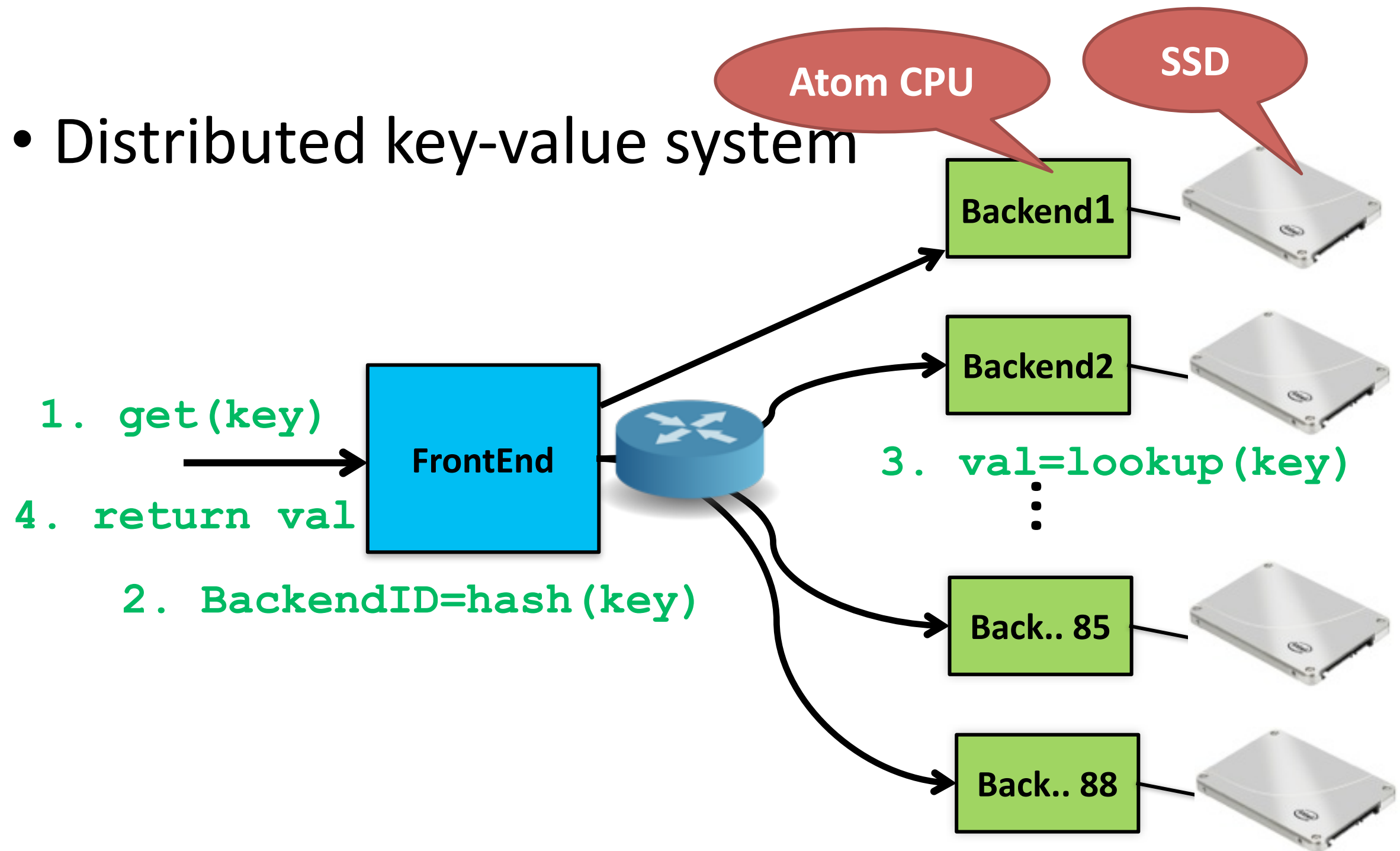
And now... Load imbalance

- Distributed key-value system



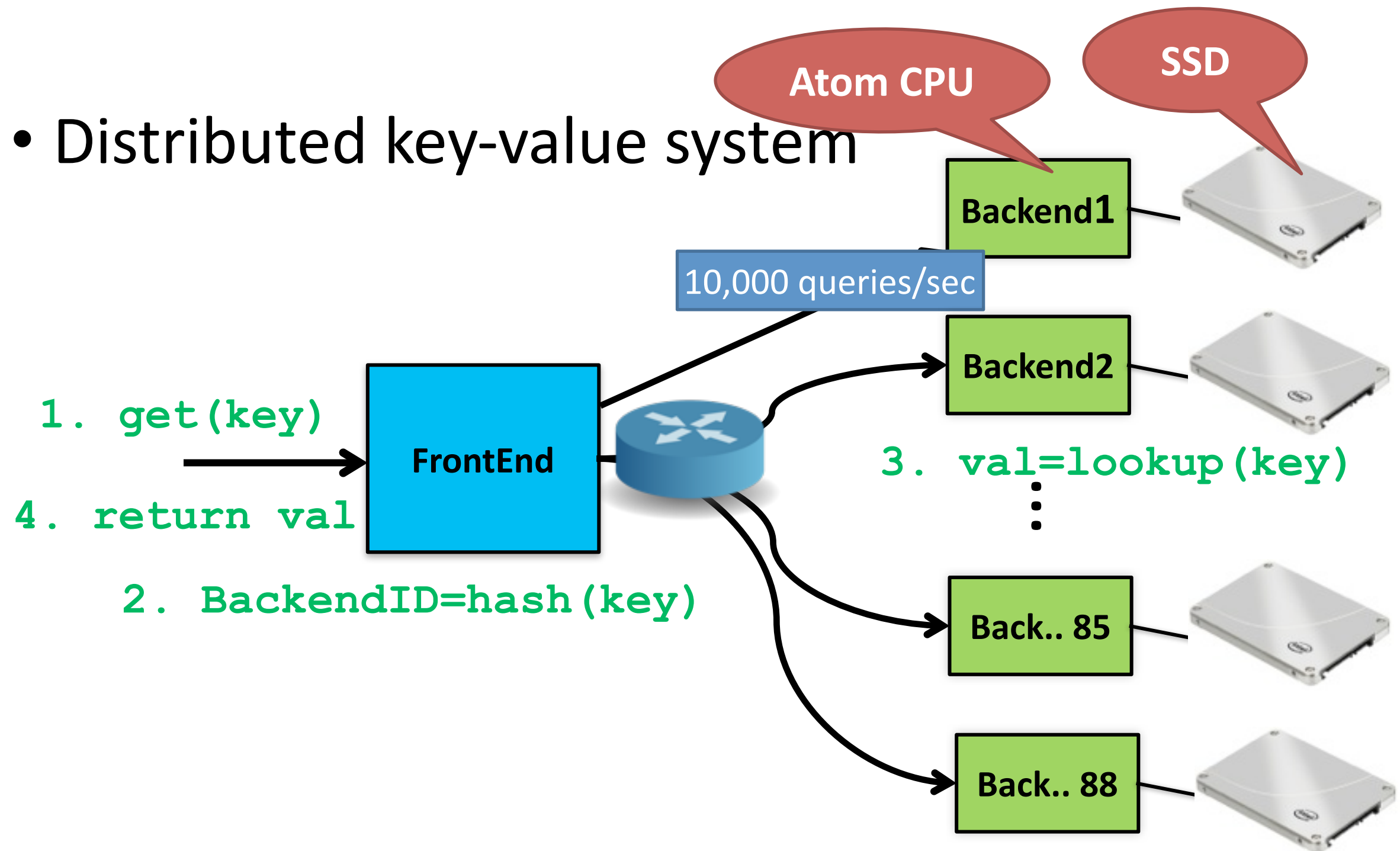
And now... Load imbalance

- Distributed key-value system



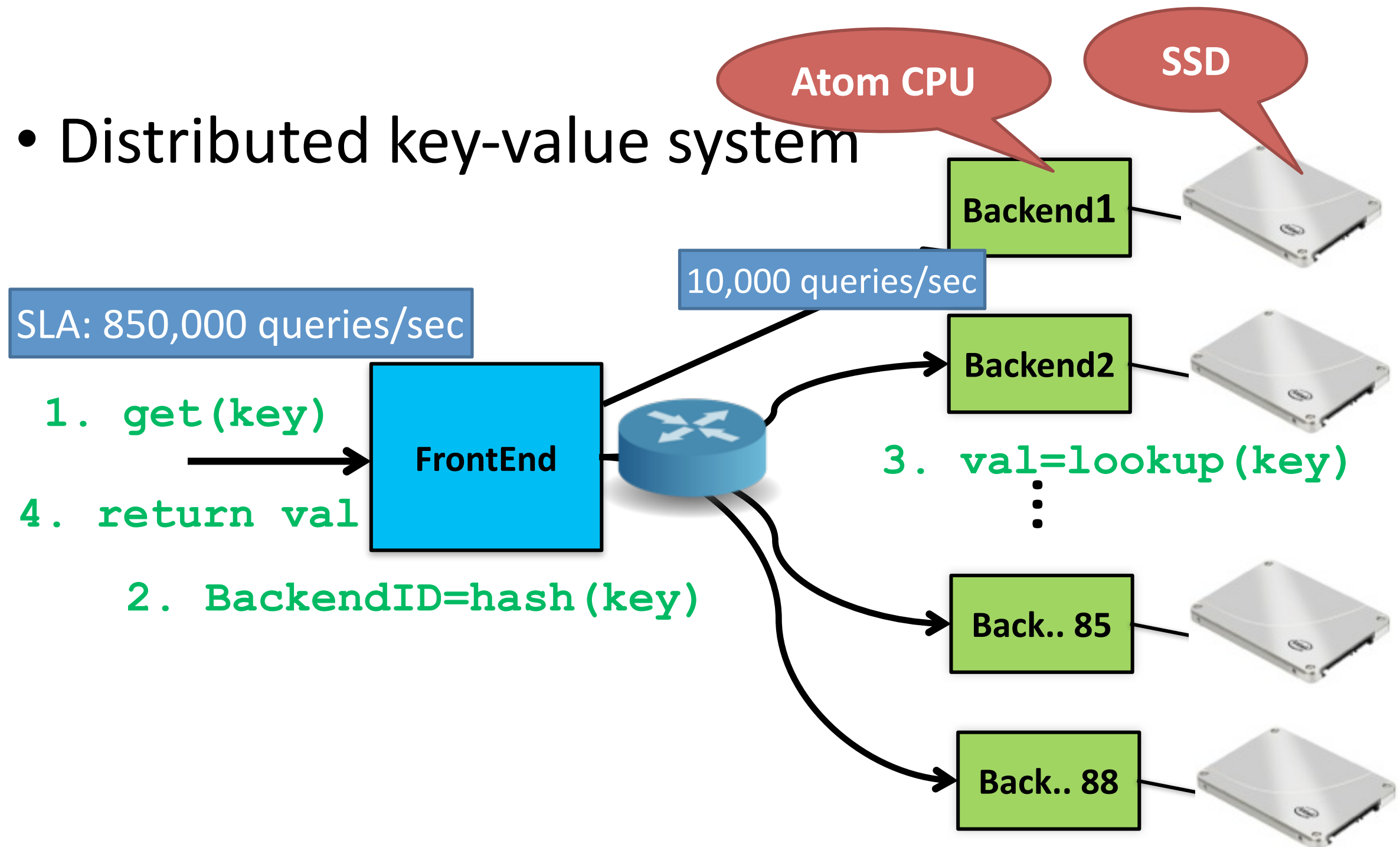
And now... Load imbalance

- Distributed key-value system

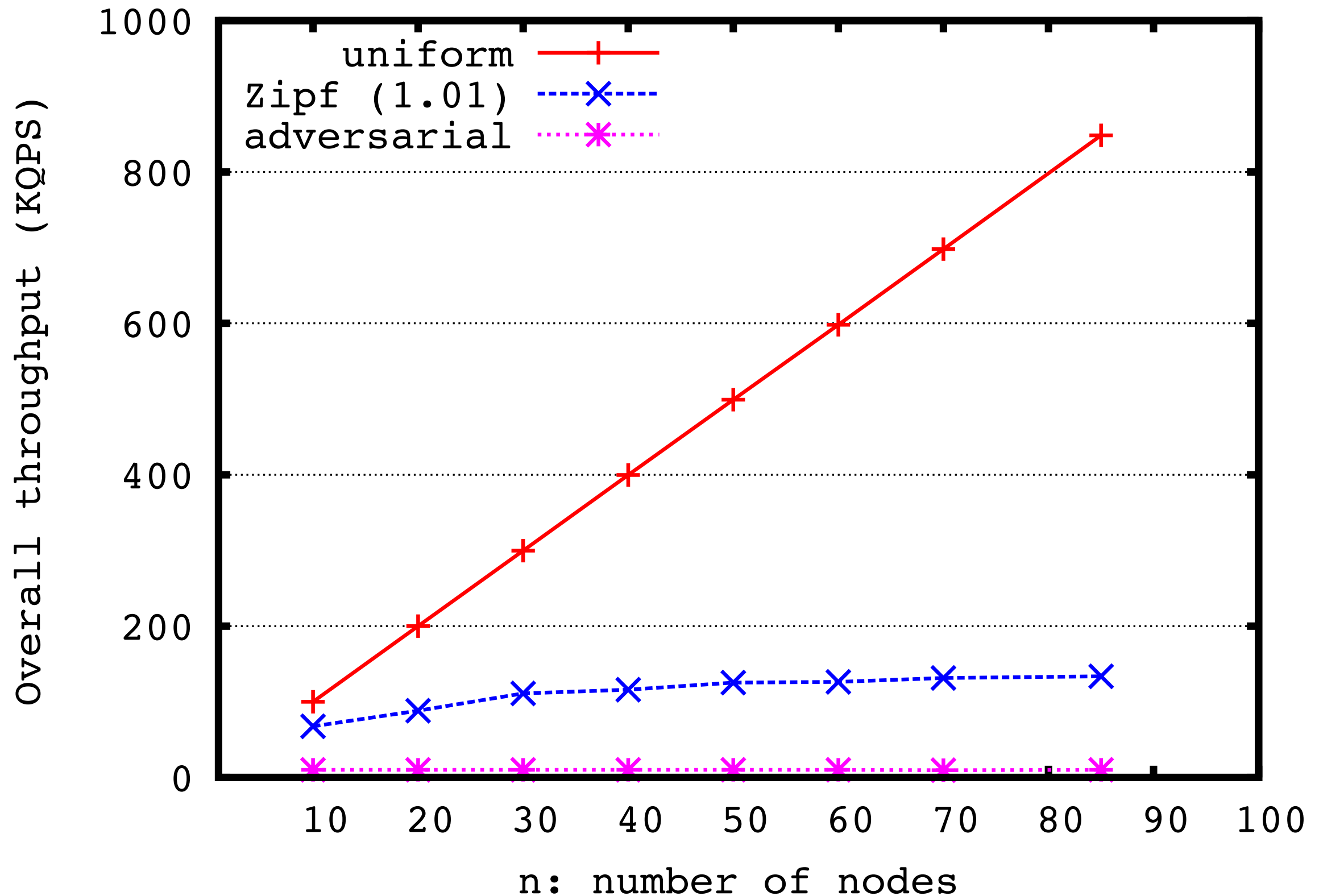


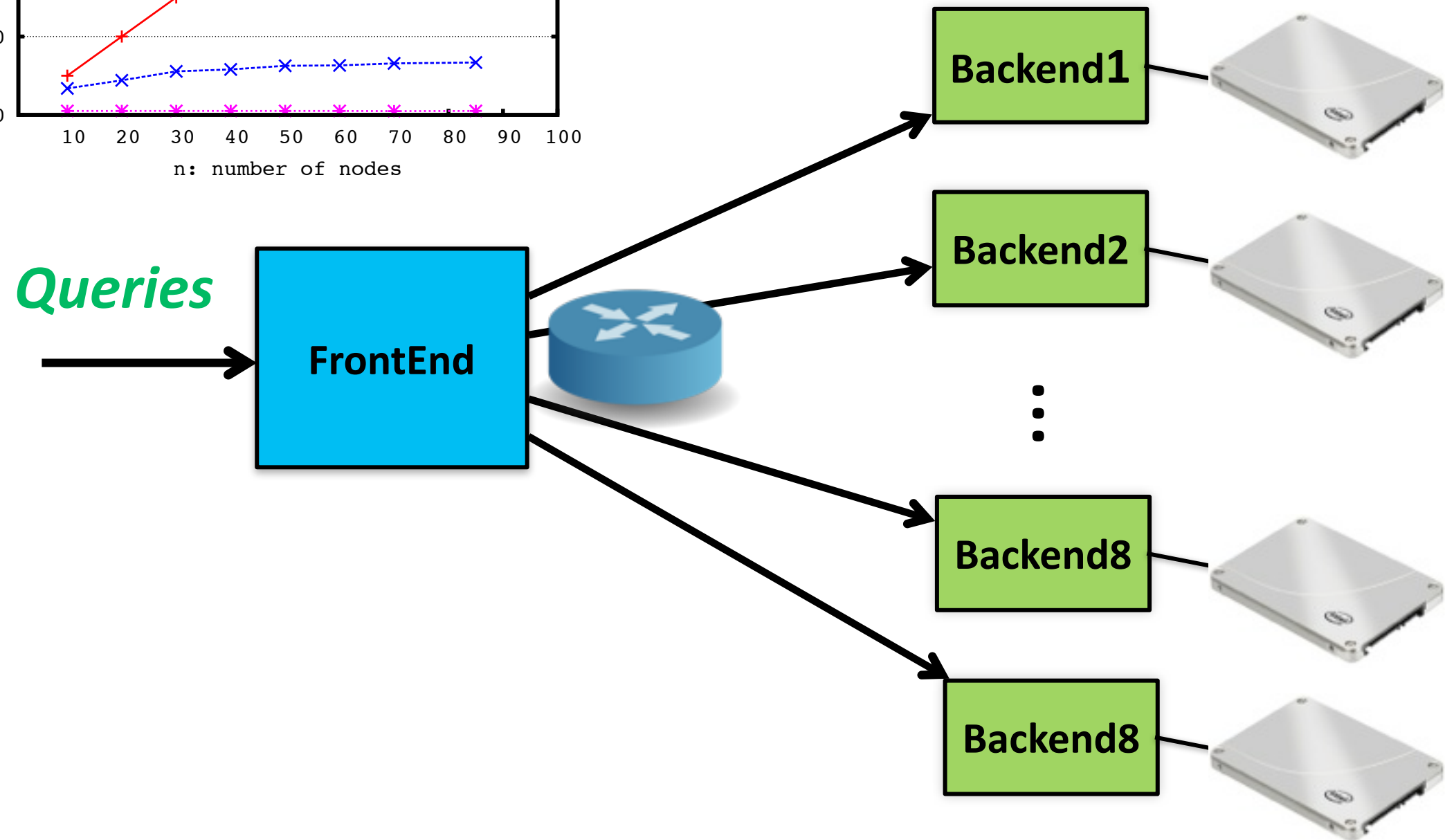
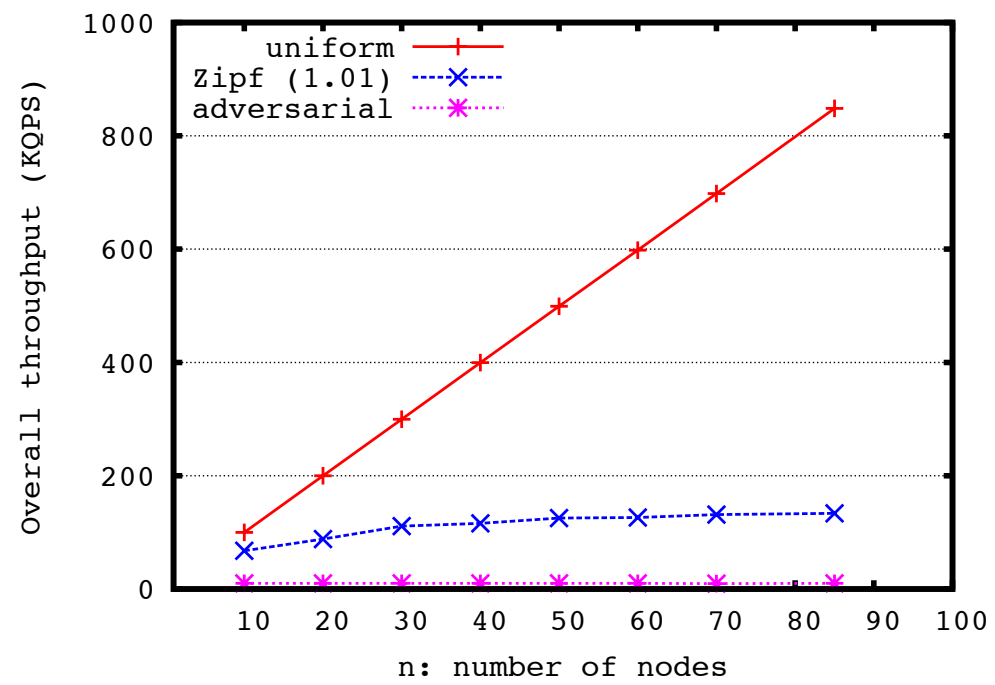
And now... Load imbalance

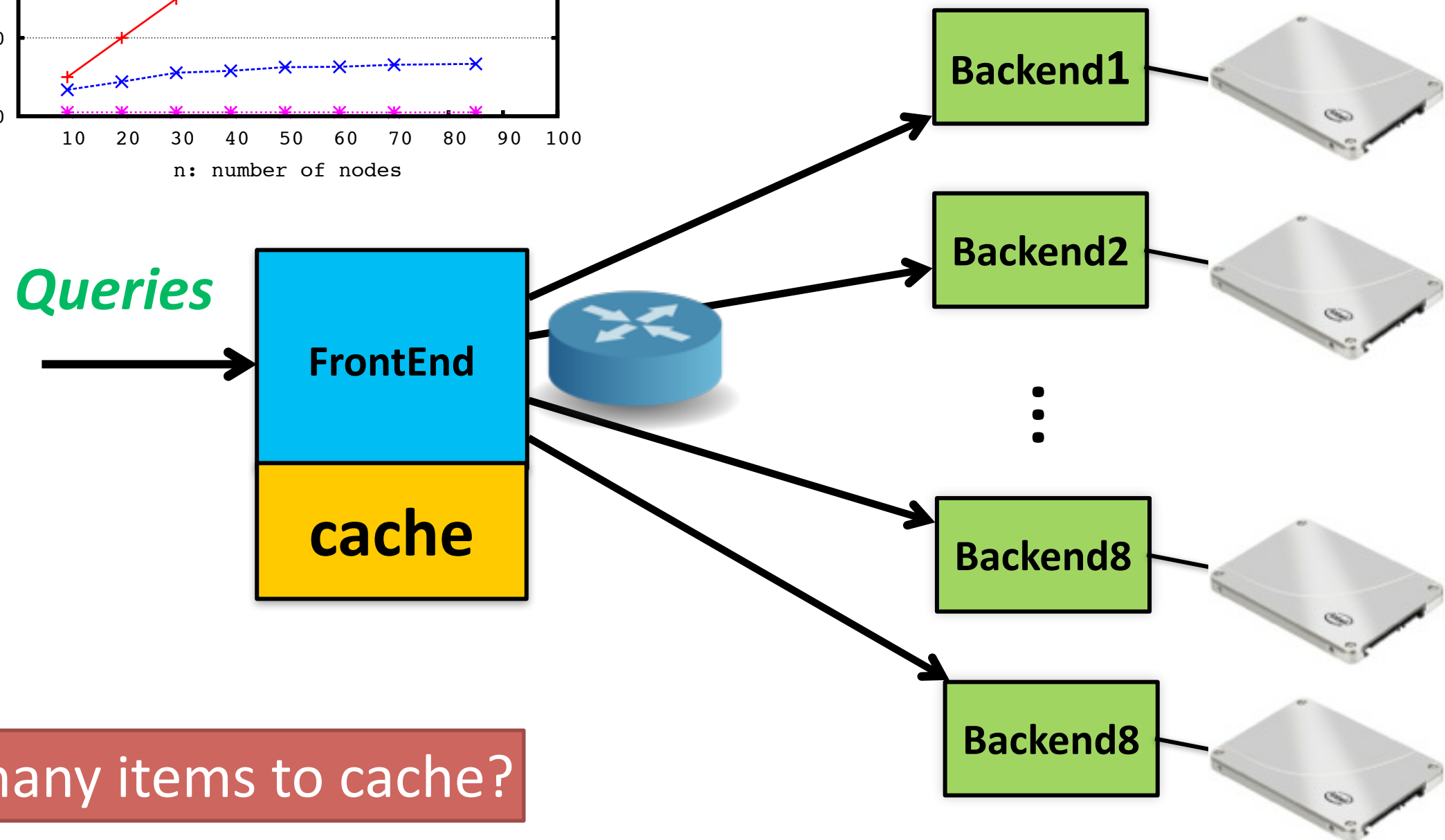
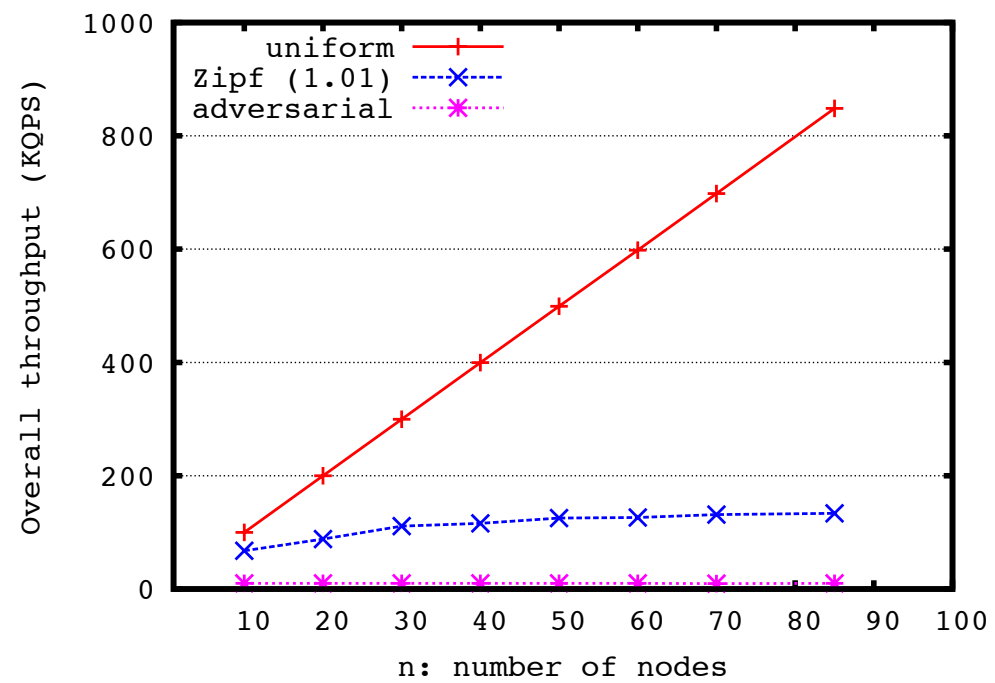
- Distributed key-value system



Measured tput on FAWN testbed

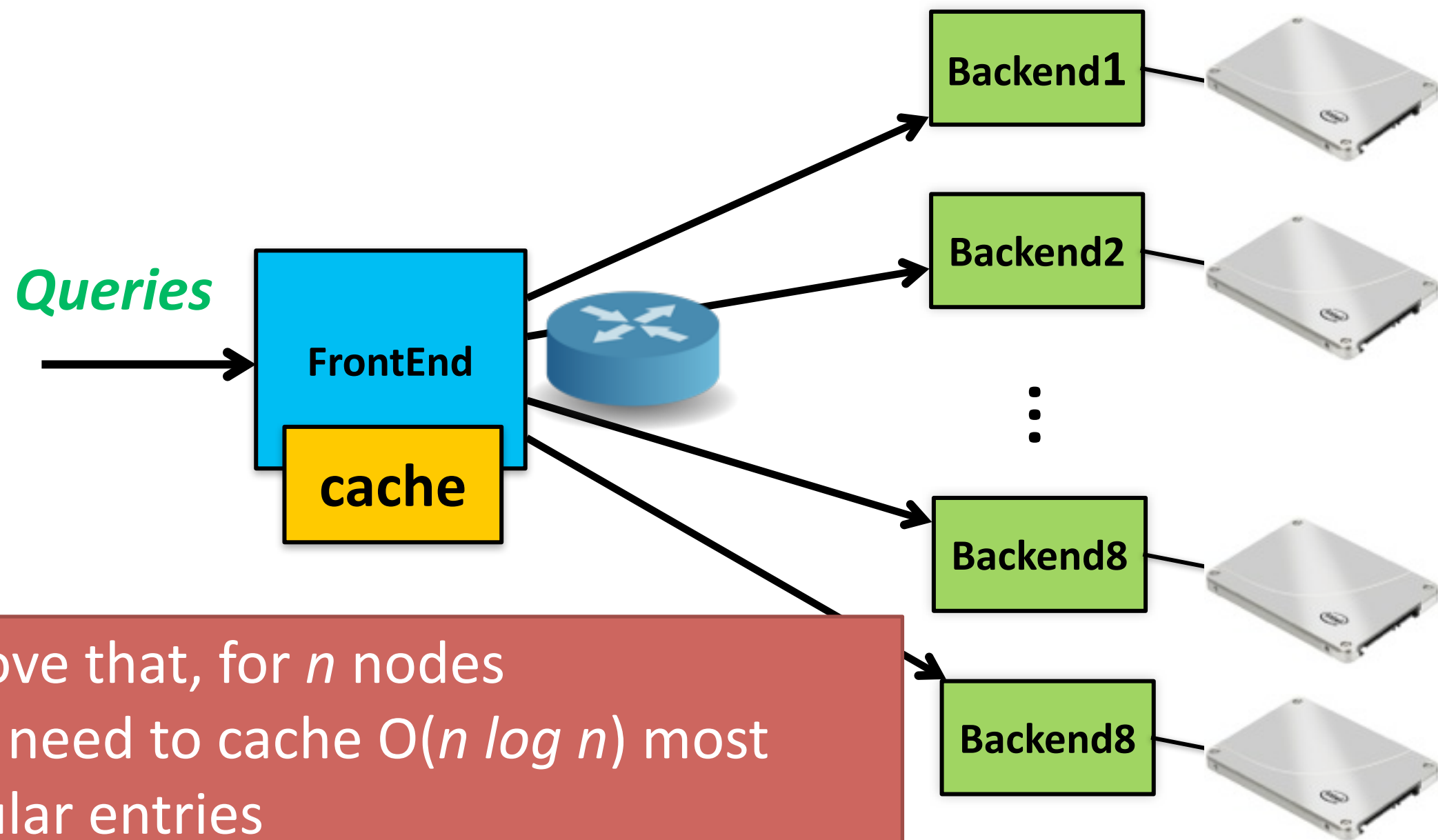






How many items to cache?

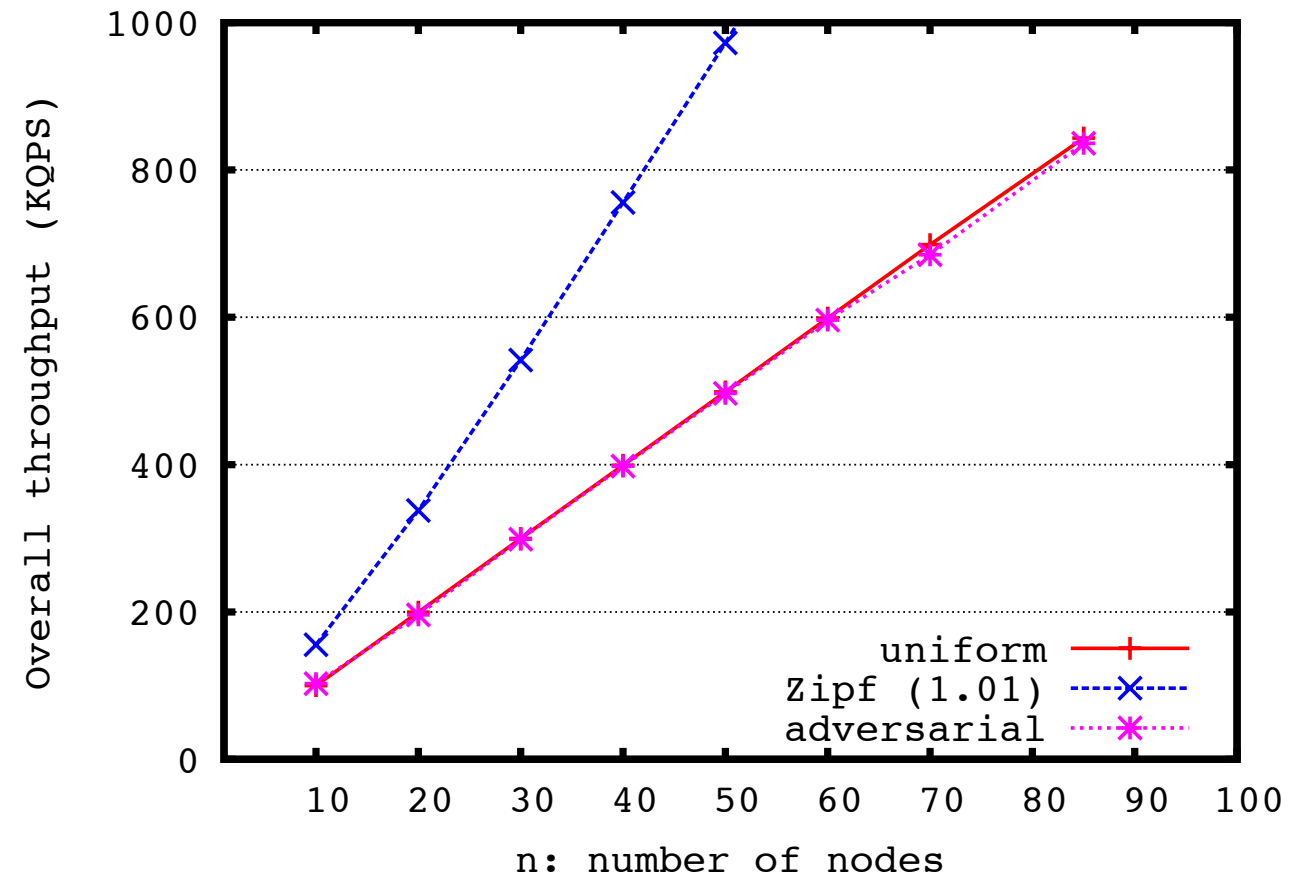
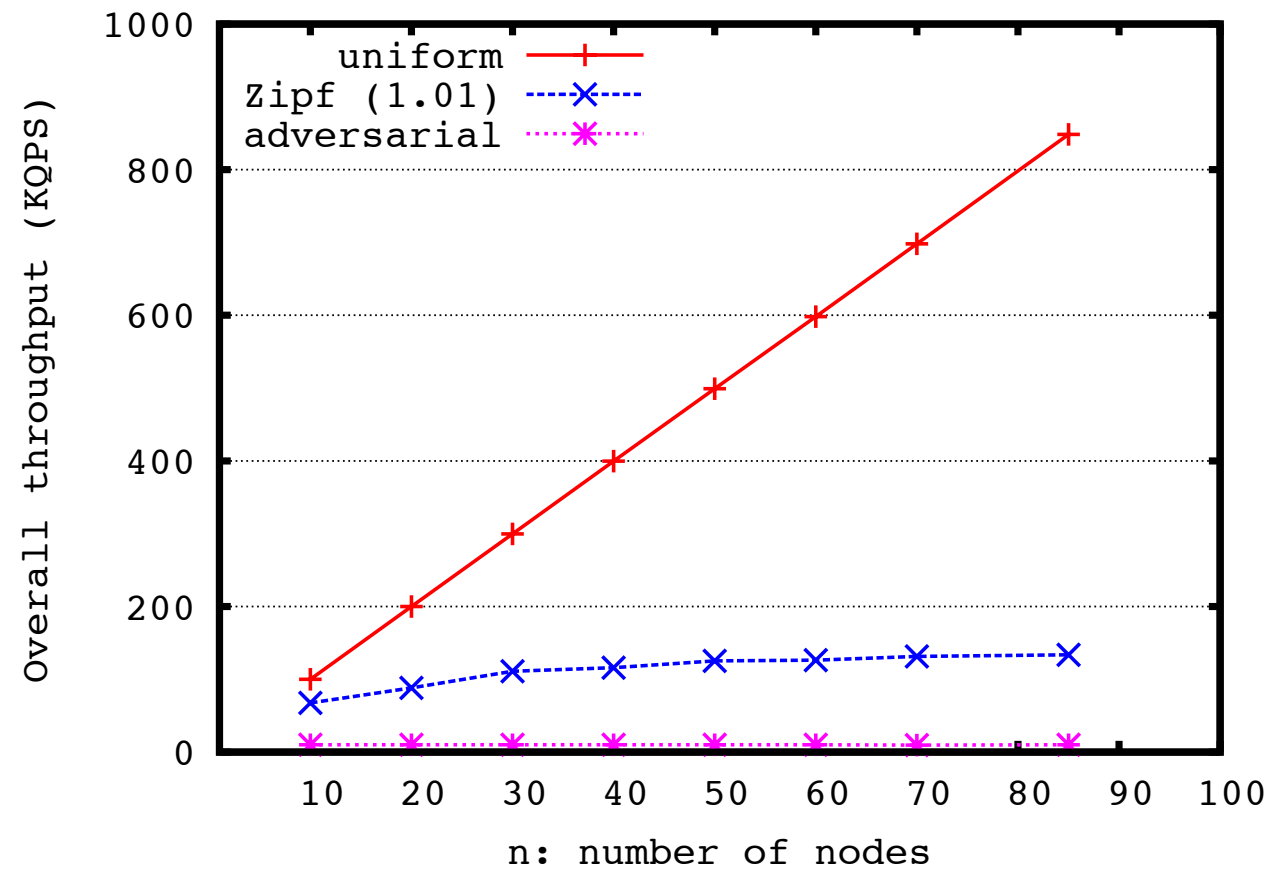
small/fast cache is enough!



We prove that, for n nodes

- Only need to cache $O(n \log n)$ most popular entries
- With 100 backend nodes, need only about 4,000 items in the cache. Tiny!

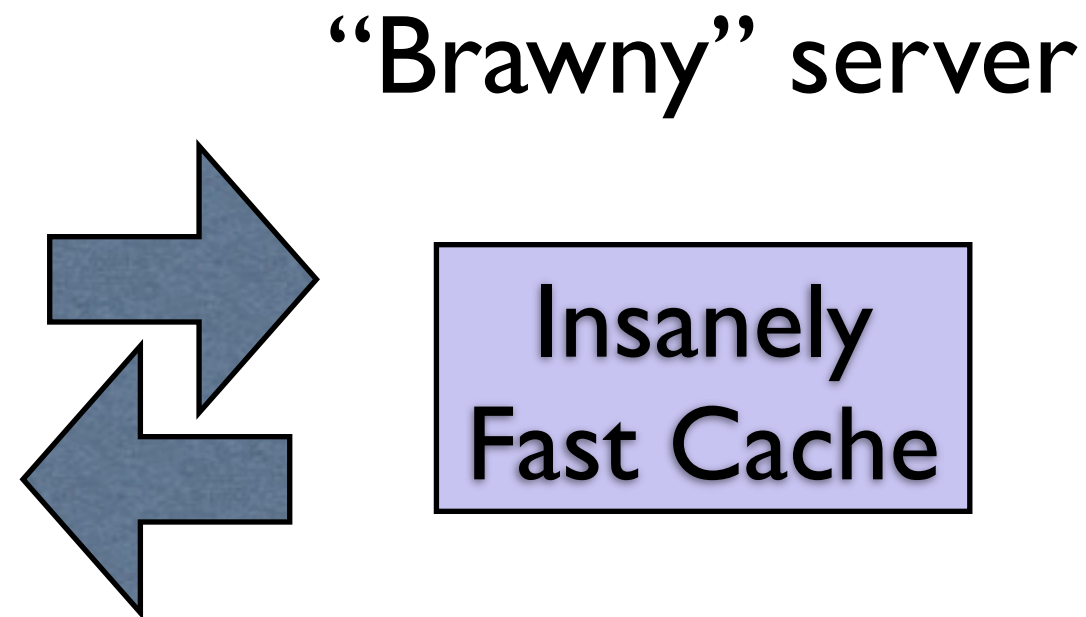
Worst case? Now best case



Thus...

FAWN-DS FAWN-KV SILT Small Cache Cuckoo

“Wimpy” servers [FAWN, SOSP 2009]



[SILT, SOSP 2011]

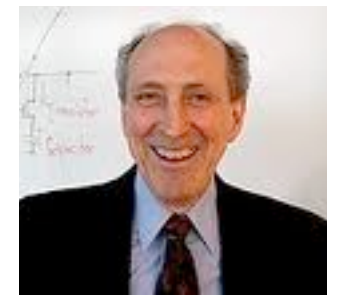
$O(N \log N)$ [“small cache” socc 2011]

Multi-reader
parallel cuckoo
hashing [“MemC3” - NSDI 2013]

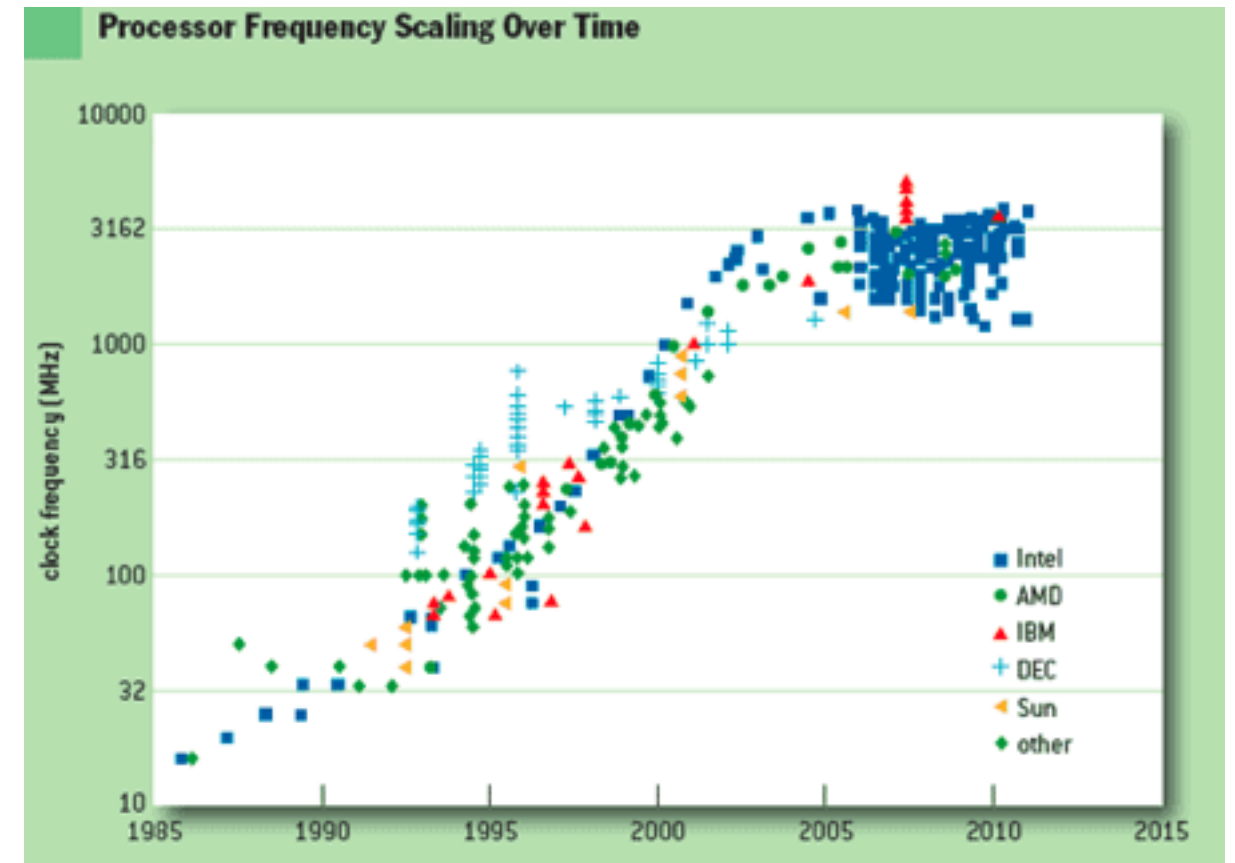
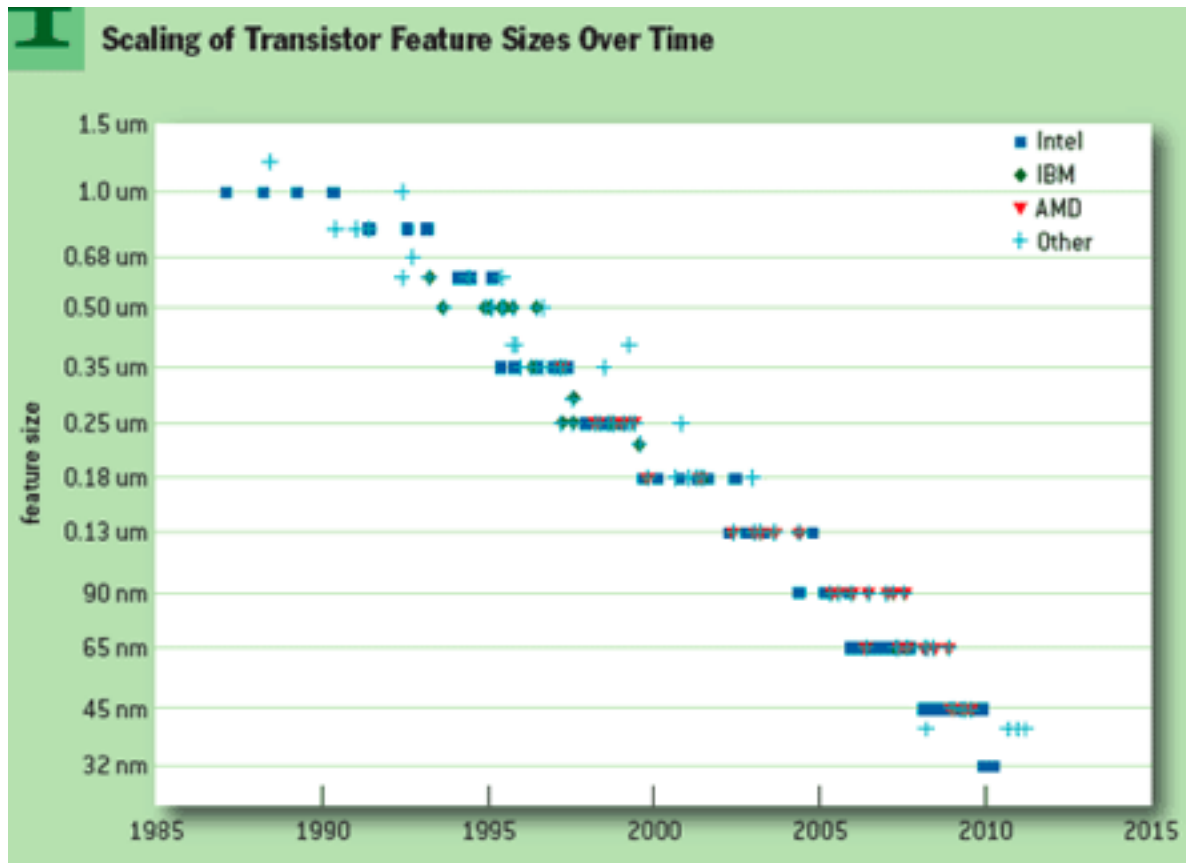
Entropy-coded tries [SOSP + ALENEX]
Partial-key cuckoo hashing
Cuckoo filter



Moore



Dennard



highly parallel, lower-GHz, (memory-constrained?):

Architectures, algorithms, and programming