Temporality

Vasily A. Sartakov
Externer Doktorant
Goslar, DE

About

- Externer Doktorant @ TU-BS
- Mostly Located in Moscow, Russia
- Leading RnD projects in area of uKernels, IDSs, Security (ksys labs)
- Engineering since 2004
  - Private and public companies, huge (Montavista) and small (startups)
- But here I am doing research as PhD student.

Persistent Systems

- Long story (GNOSIS (79), KeyKOS (80th), EROS (90th), Coyotos)
- New trigger: New memory technologies (STT-RAM, RRAM, PCRAM)
  - Density, capacity and physical size
  - Byte-addressable
  - Persistency of stored data
- New persistent systems
  - Use cases (how to use it, what is the profit?)
  - Architectures (do we still need 2nd storage?)
  - Models of persistent (What and how becomes persistent?)
  - Issues: reusing, volatile environment, etc..

Conceptions

- Language/library-based persistence
  - NV-Heaps [1]
  - Mnemosyne [2]
- Process-based persistence
  - NV-Process [3]
- System-wide
  - Whole system persistence [4]
- Hypervisor-based
  - NV-Hypervisor [5]
- File-systems (NVRAM is a storage)
  - BPFS [6], SCMFS [7], PMFS [8], Aerie [9]

Temporality::Motivation

- Cloud services
- Common failures (55%) in data centers are related to power outages [10] (2013).
- Duration:
  - Partial: 56 minutes
  - Complete unplanned: 119
- Cost: $8k per minutes
- Indirect costs:
  - Lose revenue of clients
  - Lost performance
  - Lost temporary data
  - Recovery causes high load on hardware

Motivation

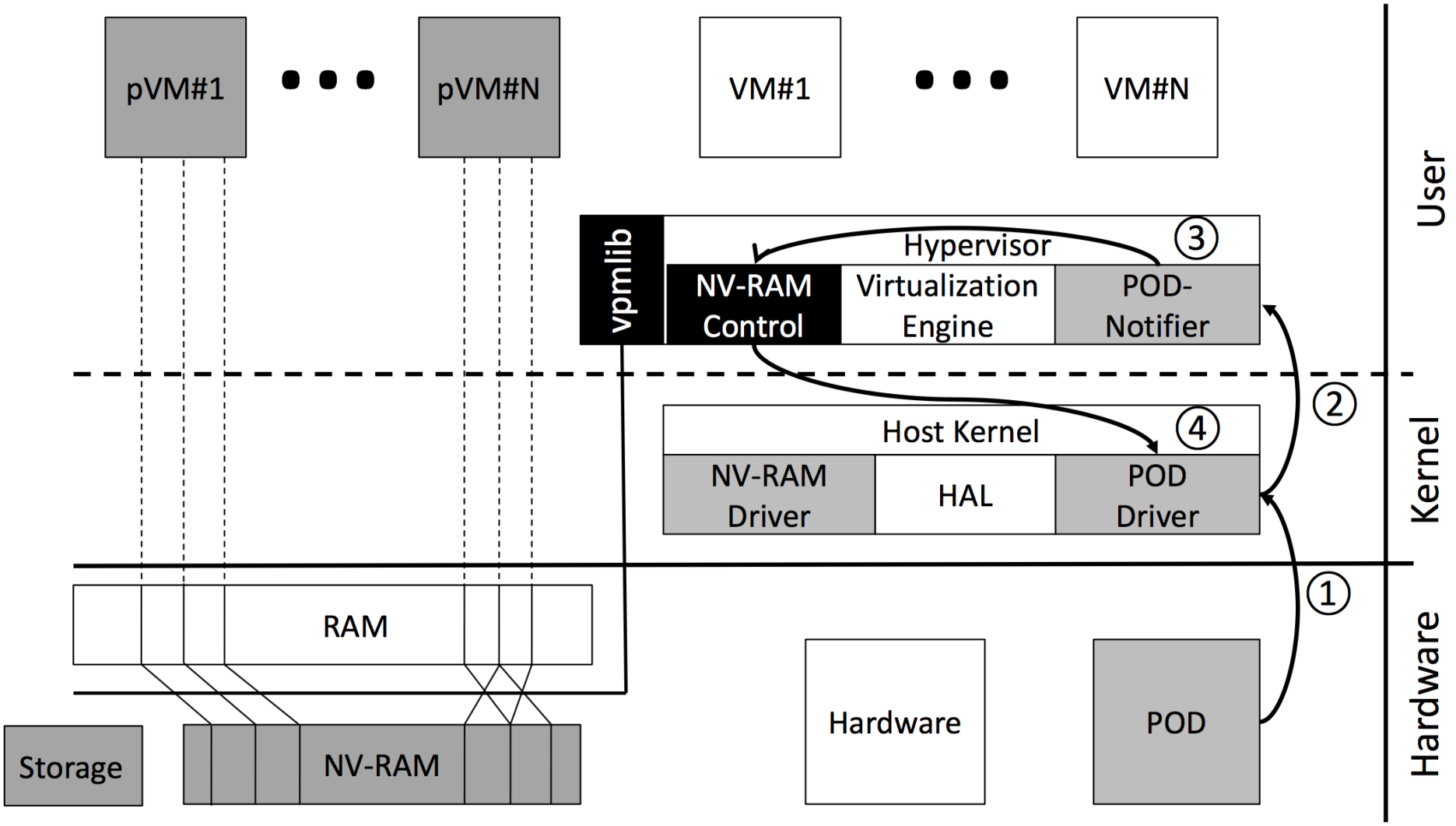
- Additional, expensive, fragile <..> hardware:
  - UPS
  - Backup power supplies

- Different way
  - NVRAM-based
  - Decrease recovery time
  - Don’t loose performance

Hardware

- Battery-backed RAM
  - Ordered and atomic writings
  - DIMMs equal performance
- Volatile CPU, Volatile devices
- Residual energy
  - Rapilog [10]
- Catch the moment of Power outage by POD

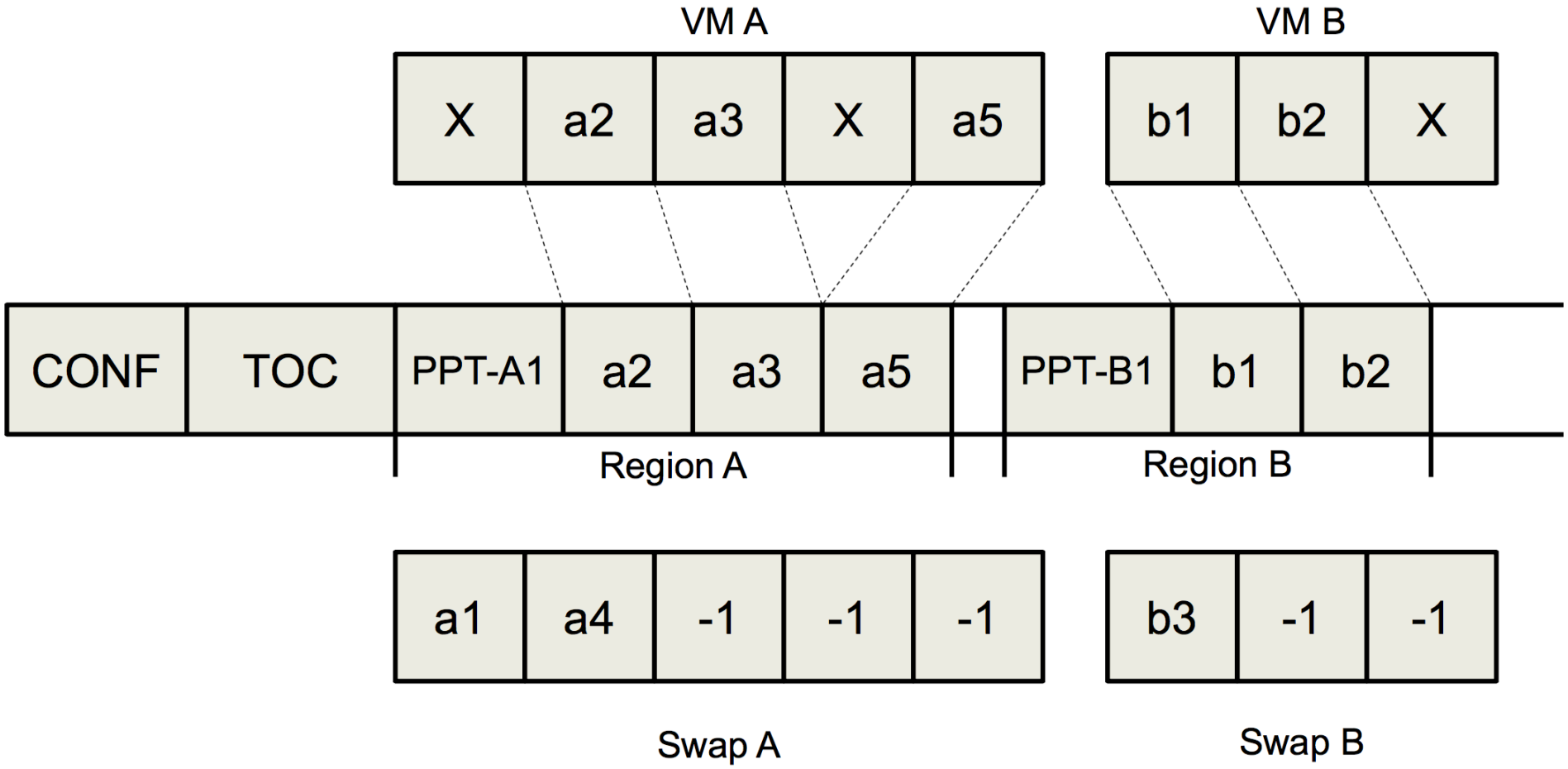
Architecture



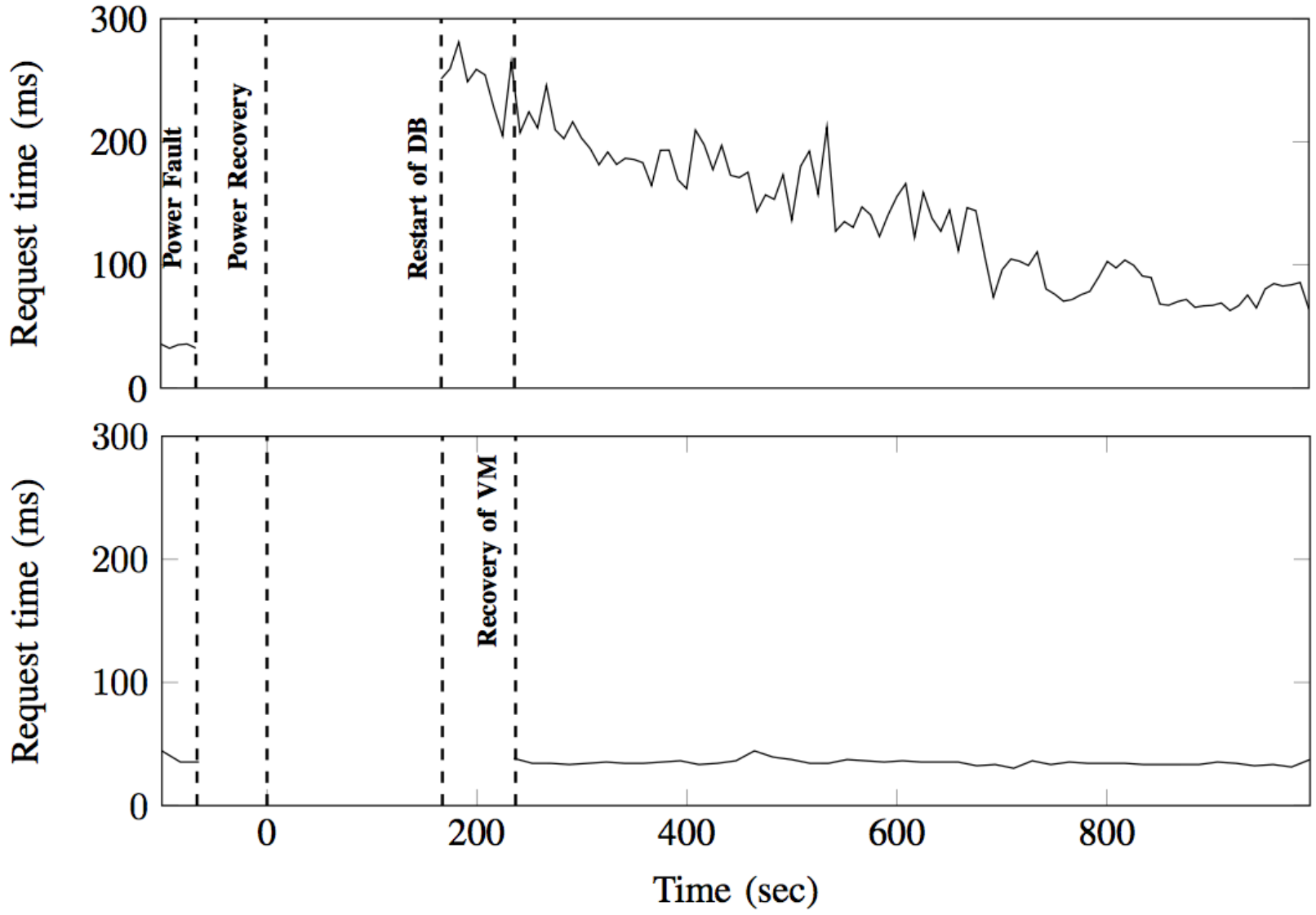
Components

- Name-based allocator
  - Control sequence of allocation/reallocation
- Virtual memory support
  - Supports multiple pVM use
- Persistent VMs:
  - Device state (should be “fixated” in a moment of power outage)
  - Memory image (Persistent memory and swap)
- Life cycle:
  - Creation, execution
  - Power outage: save the context of pVM, flush caches
  - Continue execution of pVM after power recovery

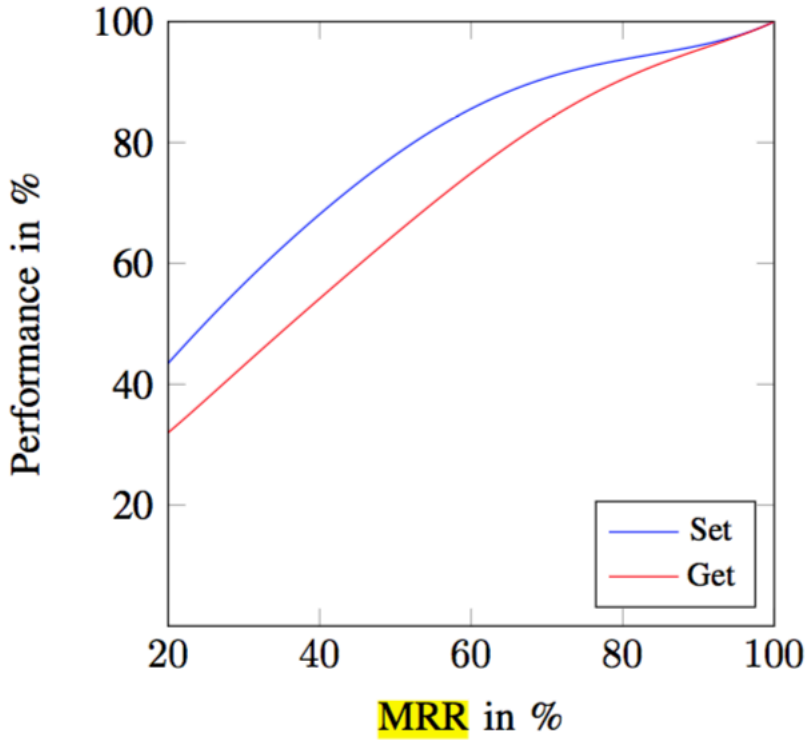
Allocator


VM A
VM B
X
a2
a3
X
a5
b1
b2
X
CONF
TOC
PPT-A1
a2
a3
a5
PPT-B1
b1
b2
Region A
Region B
a1
a4
-1
-1
-1
b3
-1
-1
Swap A
Swap B

Benchmarks

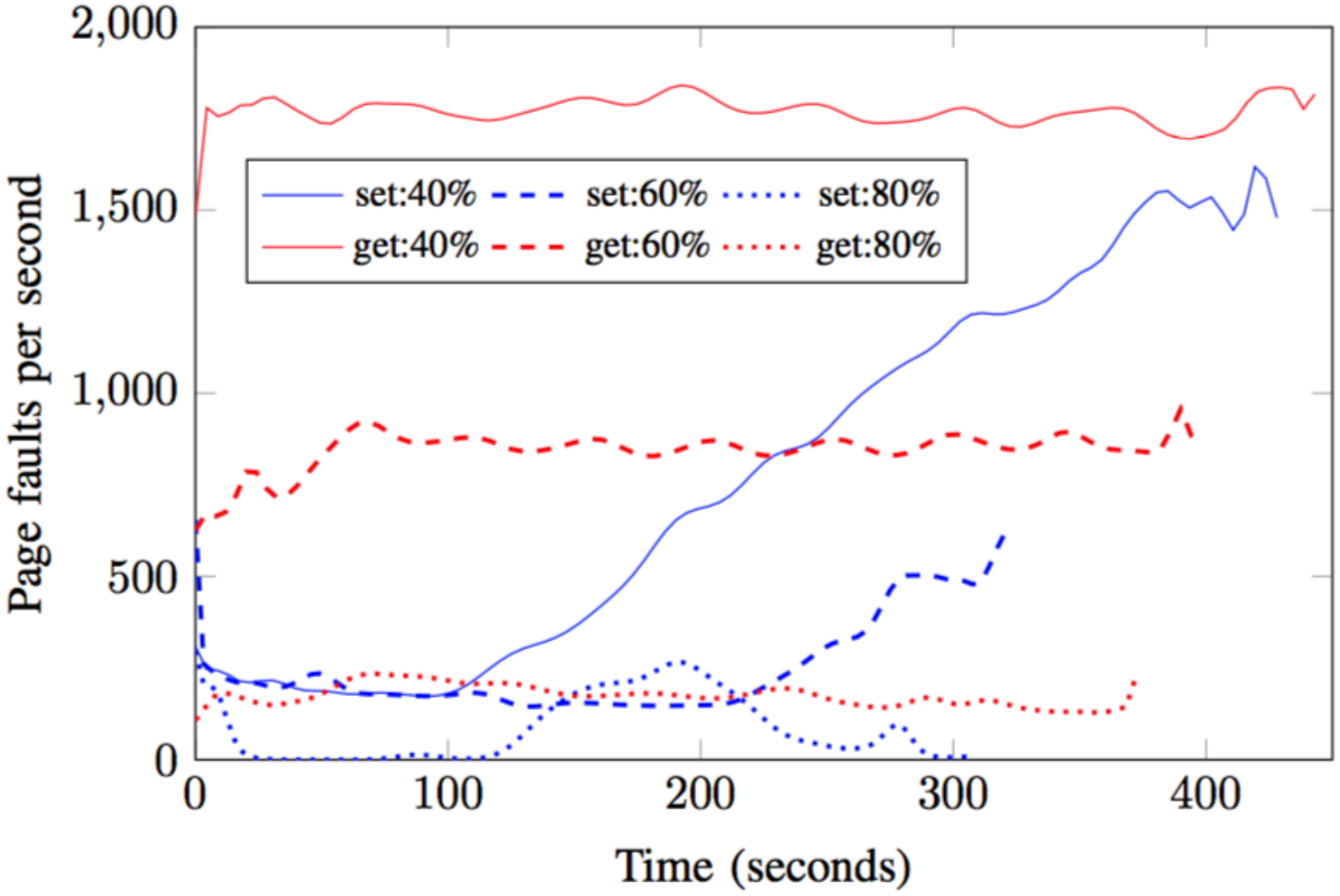

Request time (ms)
300
200
100
0
Power Fault
Power Recovery
Restart of DB
Request time (ms)
300
200
100
0
Recovery of VM
0
200
400
600
800
Time (sec)

Memory Residence Ratio



MRR = InMemory/(InMemory + InSwap)*100

Page faults per second (fixed workload)



Future work

- RAID-like protection for pVMs
  - Distributed replicas of pVM
  - RDMA as communication
  - State-less virtual environment?

References

[1]. J. Coburn, A. M. Caulfield, A. Akel, L. M. Grupp, R. K. Gupta, R. Jhala, and S. Swanson, "NV-Heaps: making persis- tent objects fast and safe with next-generation, non-volatile memories," in *ACM SIGARCH Comp. Arch. News*, 2011.

[2]. H. Volos, A. J. Tack, and M. M. Swift, "Mnemosyne: Lightweight persistent memory," in *ACM SIGARCH Computer Architecture News*, vol. 39, 2011.

[3]. X. Li, K. Lu, X. Wang, and X. Zhou, "Nv-process: A fault-tolerance process model based on non-volatile memory," in *Proceedings of the Third ACM SIGOPS Asia-Pacific Conference on Systems*, ser. APSys '12, 2012.

[4]. D. Narayanan and O. Hodson, "Whole-system persistence," in *ACM SIGARCH Computer Architecture News*, vol. 40, 2012.

[5]. V. A. Sartakov and R. Kapitza, "Nv-hypervisor: Hypervisor- based persistence for virtual machines," in *Dependable Systems and Networks (DSN)*, 2014.

[6] J. Condit, E. B. Nightingale, C. Frost, E. Ipek, B. Lee, D. Burger, and D. Coetzee, "Better I/O through byte-addressable, persistent memory," in *Proc. of the ACM SIGOPS 22nd Symposium on Operating Systems Principles*, 2009.

[7] X. Wu and A. Reddy, "Scmfs: a file system for storage class memory," in *Proc. of Int. Conf. for High Performance Computing, Networking, Storage and Analysis*, 2011.

[8] S. R. Dulloor, S. Kumar, A. Keshavamurthy, P. Lantz, D. Reddy, R. Sankaran, and J. Jackson, "System software for persistent memory," in *Proc. of the 9th European Conference on Computer Systems*, 2014.

[9] H.Volos,S.Nalli,S.Panneerselvam,V.Varadarajan,P.Saxena, and M. M. Swift, "Aerie: flexible file-system interfaces to storage-class memory," in *Proc. of the 9th European Conference on Computer Systems*, 2014.

[10] G. Heiser, E. Le Sueur, A. Danis, A. Budzynowski, T.- l. Salomie, and G. Alonso, "RapiLog: reducing system complexity through verification," in *Proc. of the 8th ACM European Conference on Computer Systems*, 2013.

[11] "2013 Cost of Data Center Outages," http://www. emersonnetworkpower.com/documents/en-us/brands/liebert/ documents/white%20papers/2013_emerson_data_center_ cost_downtime_sl-24680.pdf, Ponemon Institute, 2013.

Thank you