HotSnap: A Hot Distributed Snapshot System for Virtual Machine Cluster

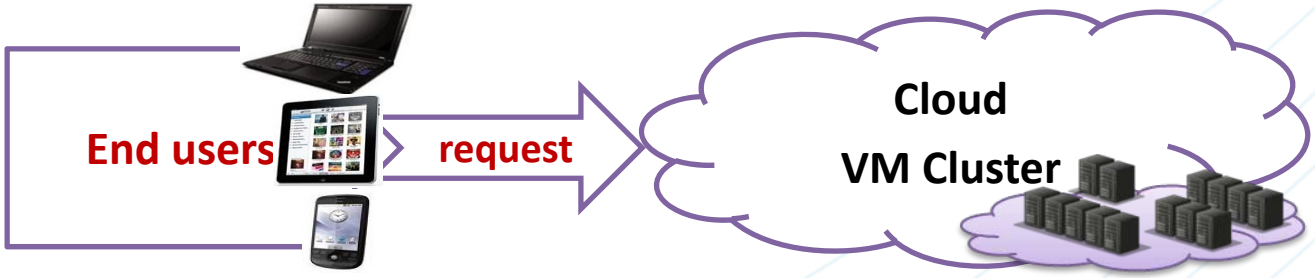
Presented by: Dr Jianxin Li
Lei Cui, Bo Li, Yangyang Zhang and Jianxin Li
ACT lab, Beihang University
2013-11-7

Outline

- Background
- Problems
- Solution & Implementation
- Experimental Results
- Conclusions

Background

- **Virtual Machine**
  - Isolation, encapsulation, multi-instance
  - Key technique supporting cloud computing
  - Limited capacity in CPU, memory, storage
- **Virtual Machine Cluster**
  - Multiple VMs are connected together to support powerful capacity
  - Scientific computing, distributed database, web service, etc



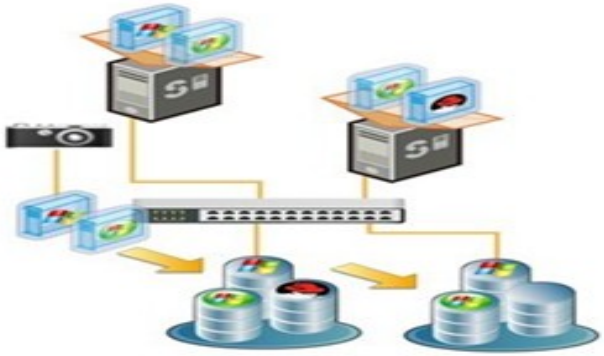
Background

- **Failure becomes a norm nowadays**
  - Computer node, Annual failure rate(AFR) is 20~60% per processor [J. Physics'07]
  - Storage node, AFR is 2%~4%, some even 3.9%~8.3% [OSDI'10]
  - Network node, AFR is 1.1%~11.4% [SIGCOMM'11]

Component	Disk	Node	Rack
MTTF	10-50years	4.3months	10.2years

Google Cluster

- **VM Skills**
  - Migration
    - **Tolerate computer node failure**
  - vLockstep
    - **Tolerate computer node failure**
  - Snapshot
    - **Tolerate computer node and software failure**



snapshot

Background

- **Failure becomes a norm nowadays**
  - Computer node, Annual failure rate(AFR) is 20~60% per processor [J. Physics'07]
  - Storage node, AFR is 2%~4%, some even 3.9%~8.3% [OSDI'10]
  - Network node, AFR is 1.1%~11.4% [SIGCOMM'11]

Component	Disk	Node	Rack
MTTF	10-50years	4.3months	10.2years

Google Cluster

- **VM Skills**
  - Migration
    - **Tolerate computer node failure**

VMC snapshot and rollback occurs frequently to survive from the failures to complete the long time task running in VMC

  - Snapshot
    - **Tolerate computer node and software failure**

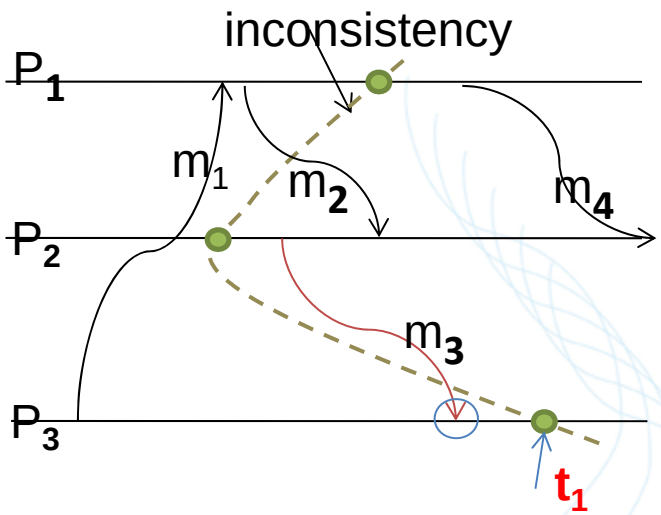
snapshot

Outline

- Background
- Problems
- Solution & Implementation
- Experimental Results
- Conclusions

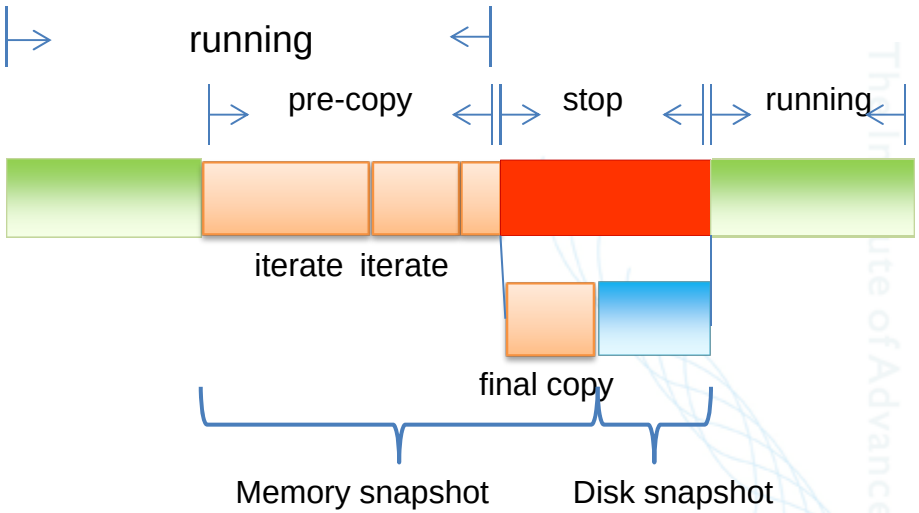
Problems

- **VMC snapshot**
  - Single VM snapshot
    - **Save memory, disk, CPU and other devices' state.**
  - Consistency protocol
    - **Global virtual time**
    - **m_3 is sent from p_2 after t_1 to p_3 before t_1, violating consistency**
    - **m_3 should be dropped to keep consistency, but this will lead to TCP-backoff**

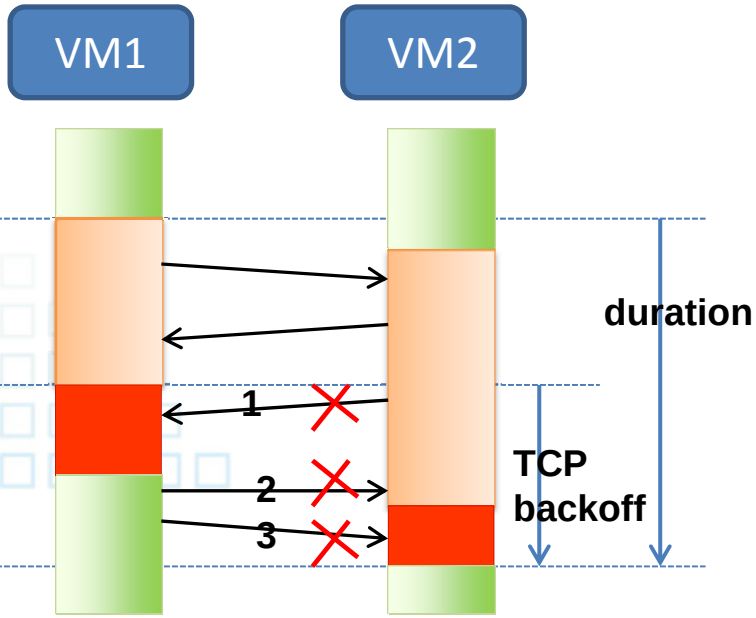


Problems

- **Analysis**
  - Current snapshot skill
    - **Stop-and-copy**
    - **Pre-copy**



Pre-copy based single VM snapshot

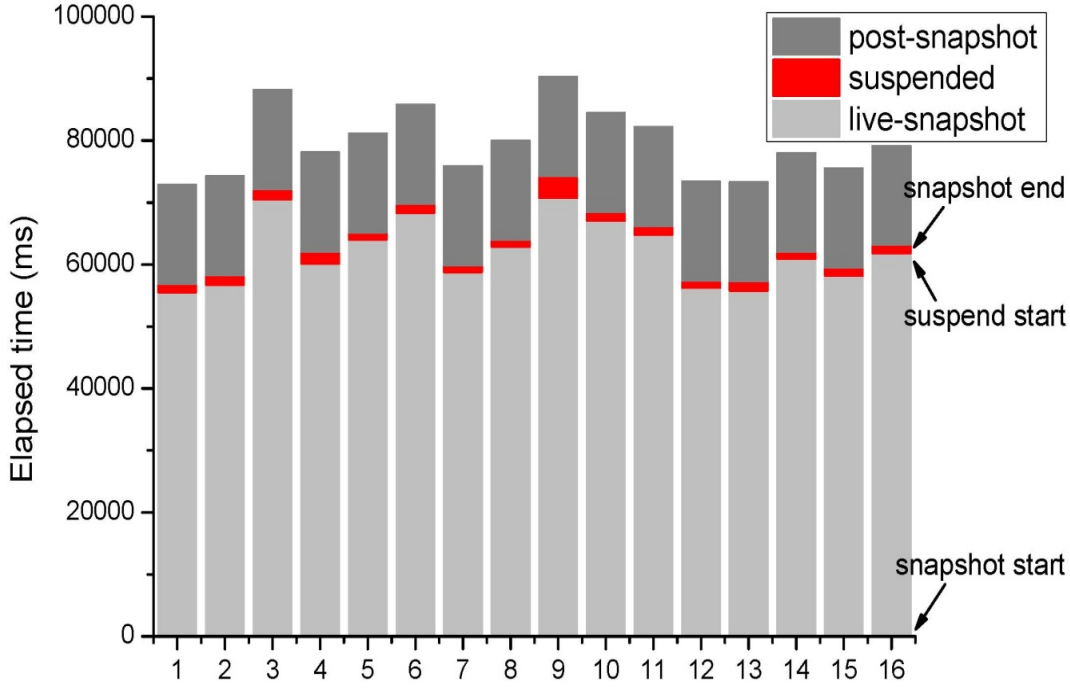


Pre-copy based VMC snapshot

TCP-backoff duration is related to downtime and difference between VMs' snapshot completion times

Problems

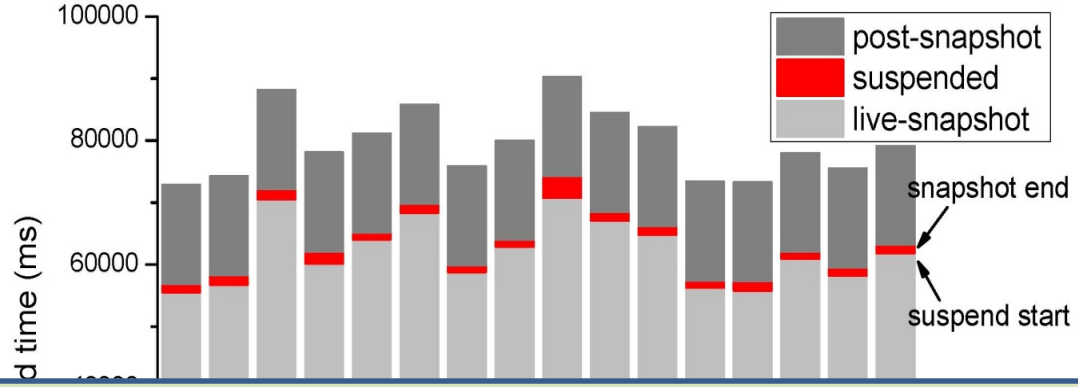
- **Experimental result, a sample**
  - 16 2G memory VMs. Distcc to compile the Linux kernel 2.6.32-5
  - VM_0 is Distcc client, TCP-backoff of VM_0 and VM_1 is 12.7s



Details of Pre-copy based VMC snapshot

Problems

- **Experimental result, a sample**
  - 16 2G memory VMs. Distcc to compile the Linux kernel 2.6.32-5
  - VM_0 is Distcc client, TCP-backoff of VM_0 and VM_1 is 12.7s



Minimize the downtime of single VM snapshot

Minimize the difference of snapshot completion times between communicating VMs

Details of Pre-copy based VMC snapshot

Outline

- Background
- Problems
- Solution & Implementation
- Experimental Results
- Conclusions

Solutions

- Key Issues in Pre-copy method
  - The downtime of single VM snapshot
    - Index tree in disk snapshot, several seconds
    - Final copy in memory snapshot, hundreds to thousands milliseconds
  - The difference of snapshot completion times
    - Iterate copy the memory state. Workload impact.
    - Available IO bandwidth to save memory state
- HotSnap solutions
  - Single VM snapshot
    - Copy-on-write(COW) based memory snapshot
    - Redirect-on-write (ROW) based disk snapshot
  - Consistency protocol
    - Suitable to virtualized environments, keep global consistency

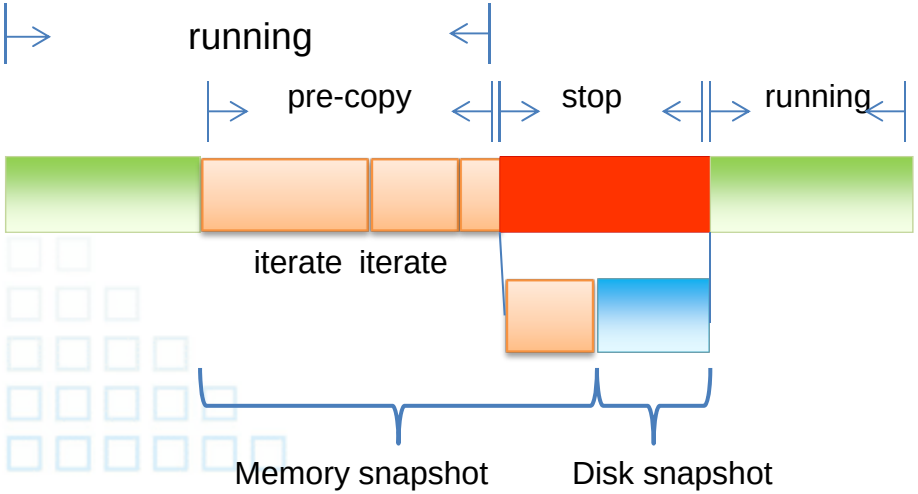
Pre-copy vs. HotSnap

Pre-copy method

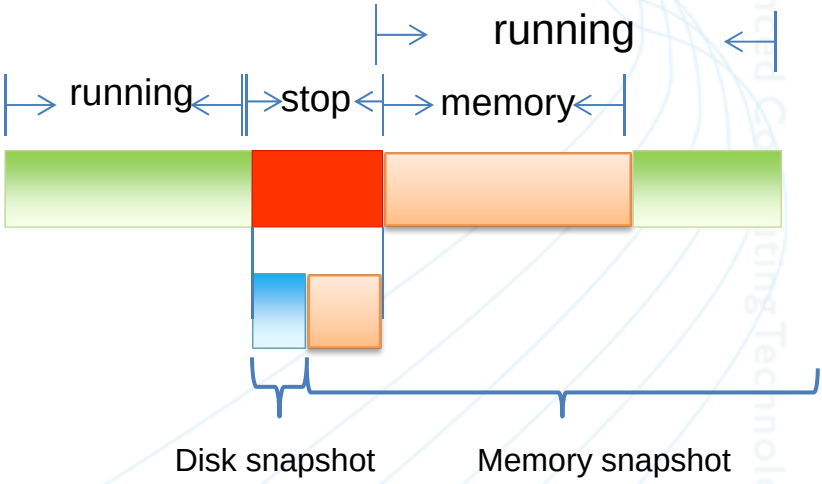
- Stop VM at the end
- Iterate copy the memory state
- Copy-on-write disk snapshot
- Longer downtime and duration

HotSnap method

- **Stop VM first**
- **Copy-on-write memory snapshot**
- **Redirect-on-write disk snapshot**
- **Short downtime and duration**



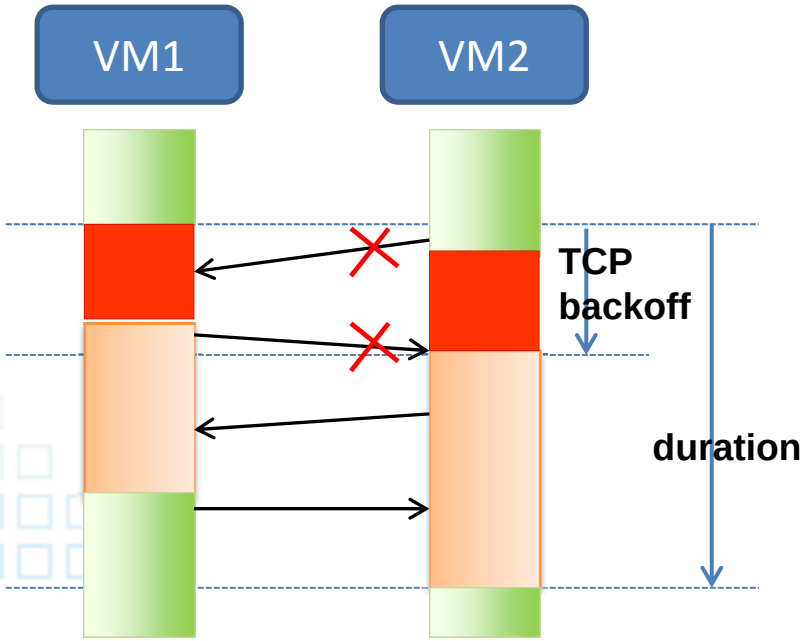
Pre-copy based single VM snapshot



HotSnap for single VM

Solutions

- **HotSnap analysis**
  - Stop VM first, record some metadata, write-protect the memory, resume the VM, and save the state during execution
  - Light-weight operations

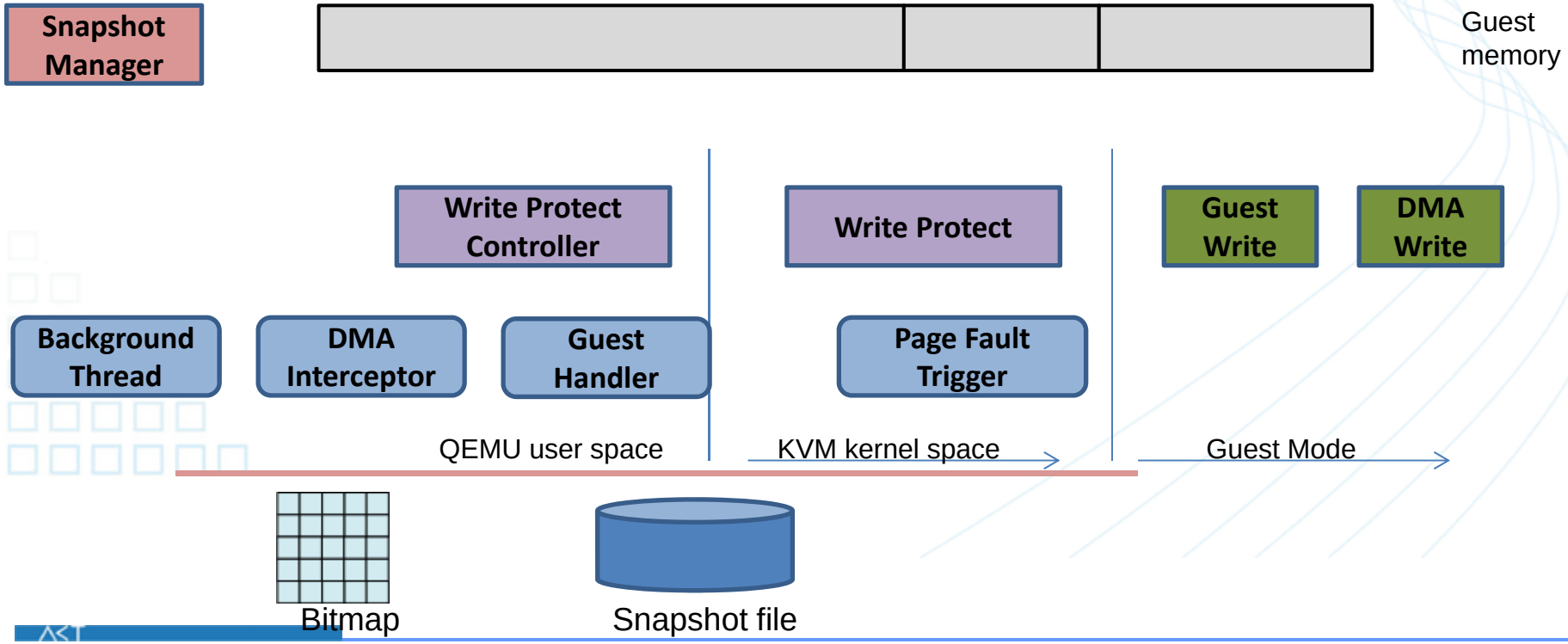


HotSnap based VMC snapshot

TCP-backoff duration is mainly related to downtime

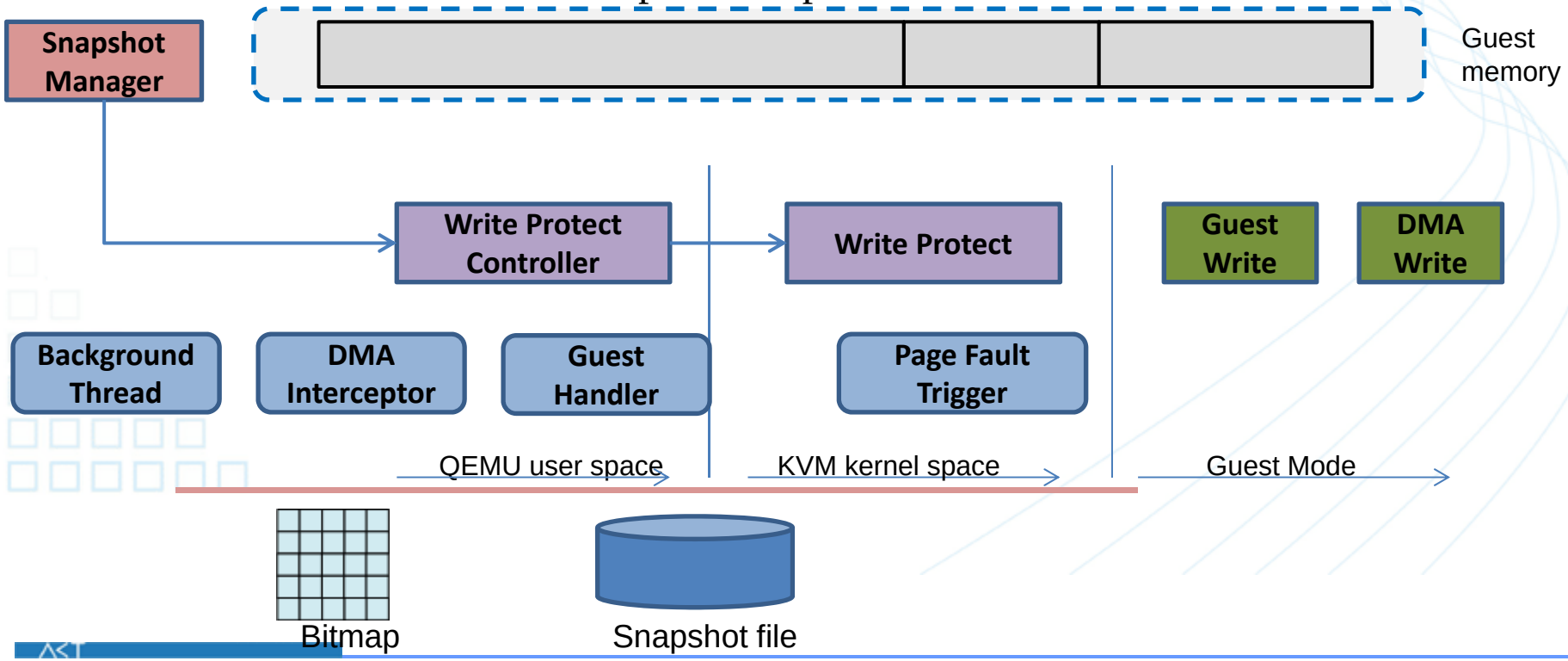
HotSnap skills – memory snapshot

- **COW based memory snapshot**
  - Only involve saving CPU, network, and devices' state, and write-protecting the memory
  - Intercept DMA write operations
  - Uniform view for bitmap and snapshot file



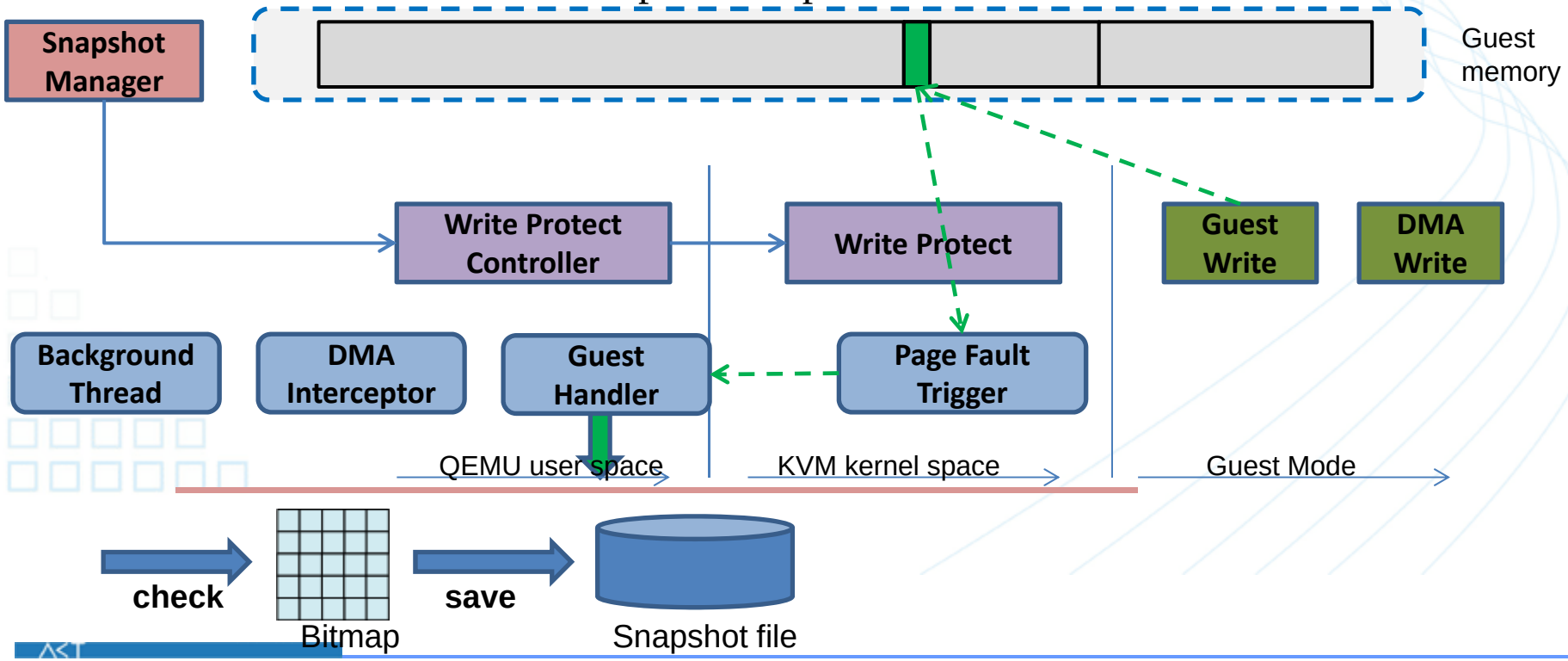
HotSnap skills – memory snapshot

- **COW based memory snapshot**
  - Only involve saving CPU, network, and devices' state, and write-protecting the memory
  - Intercept DMA write operations
  - Uniform view for bitmap and snapshot file



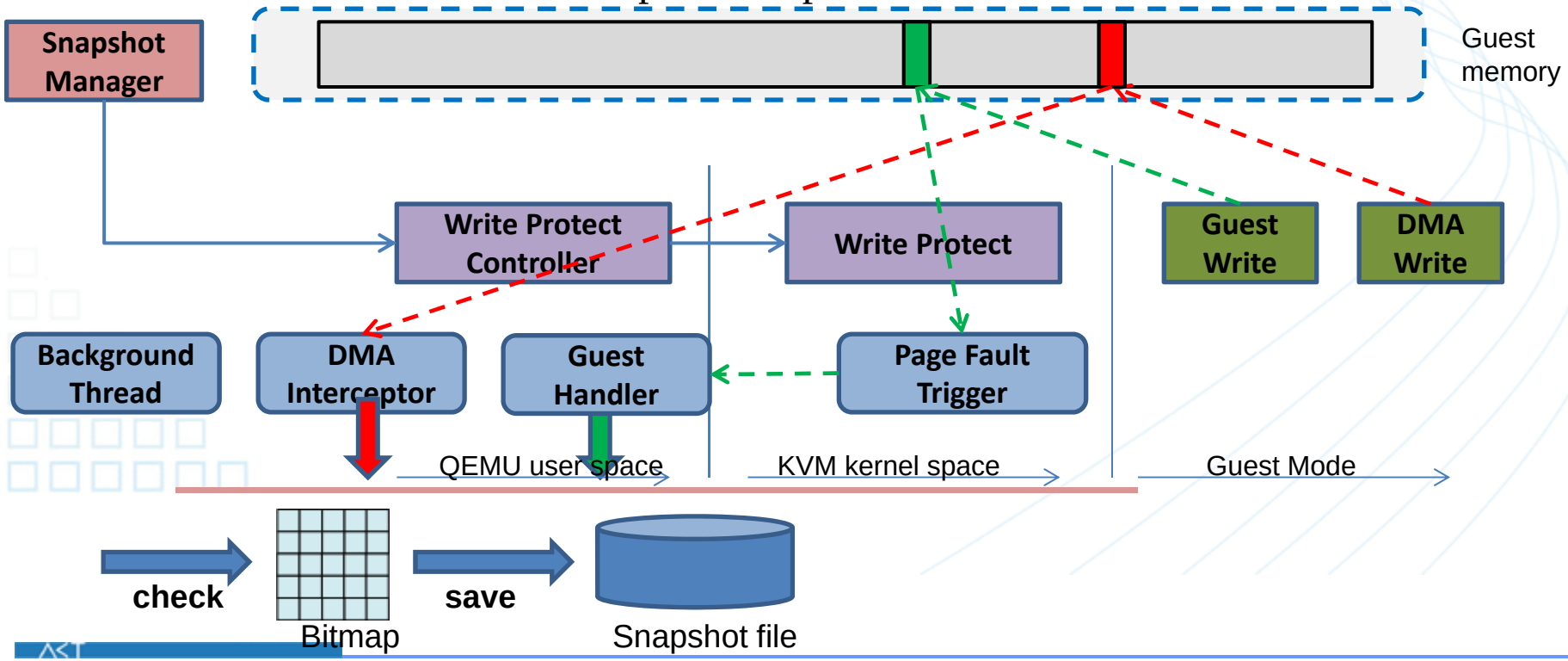
HotSnap skills – memory snapshot

- **COW based memory snapshot**
  - Only involve saving CPU, network, and devices' state, and write-protecting the memory
  - Intercept DMA write operations
  - Uniform view for bitmap and snapshot file



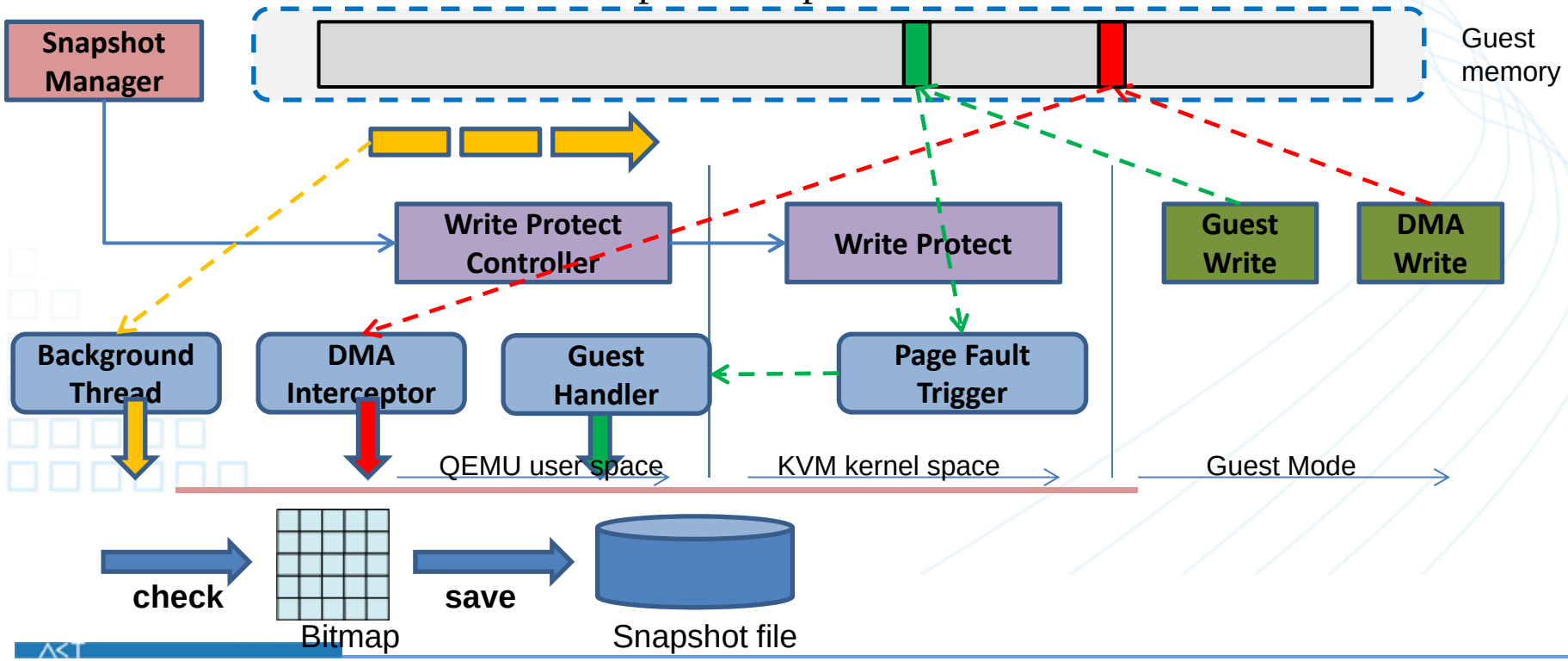
HotSnap skills – memory snapshot

- **COW based memory snapshot**
  - Only involve saving CPU, network, and devices' state, and write-protecting the memory
  - Intercept DMA write operations
  - Uniform view for bitmap and snapshot file



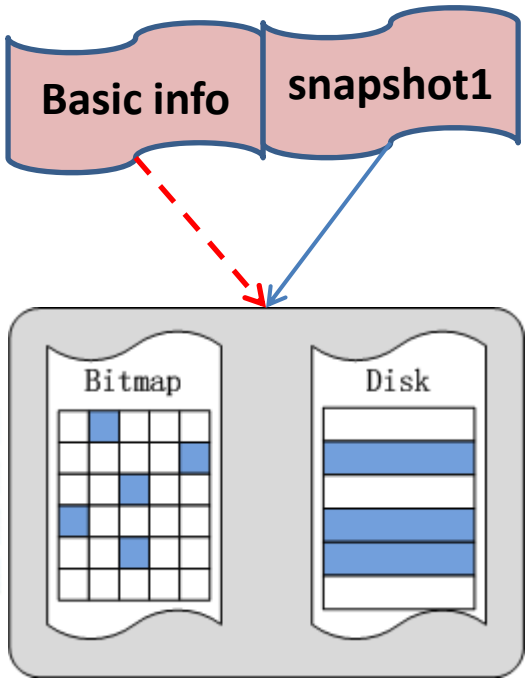
HotSnap skills – memory snapshot

- **COW based memory snapshot**
  - Only involve saving CPU, network, and devices' state, and write-protecting the memory
  - Intercept DMA write operations
  - Uniform view for bitmap and snapshot file



HotSnap skills – disk snapshot

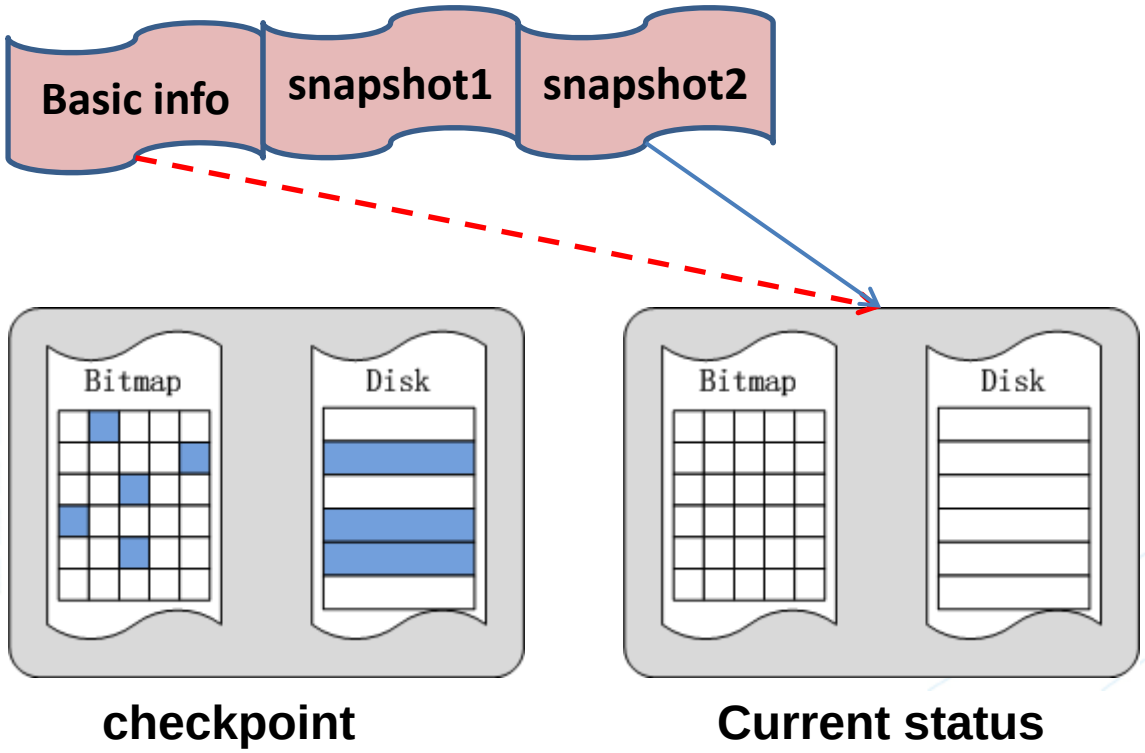
- **ROW based disk snapshot**
  - Only create one bitmap and one null disk image
  - Lightweight metadata based on bitmap
  - Redirect on write



Current status

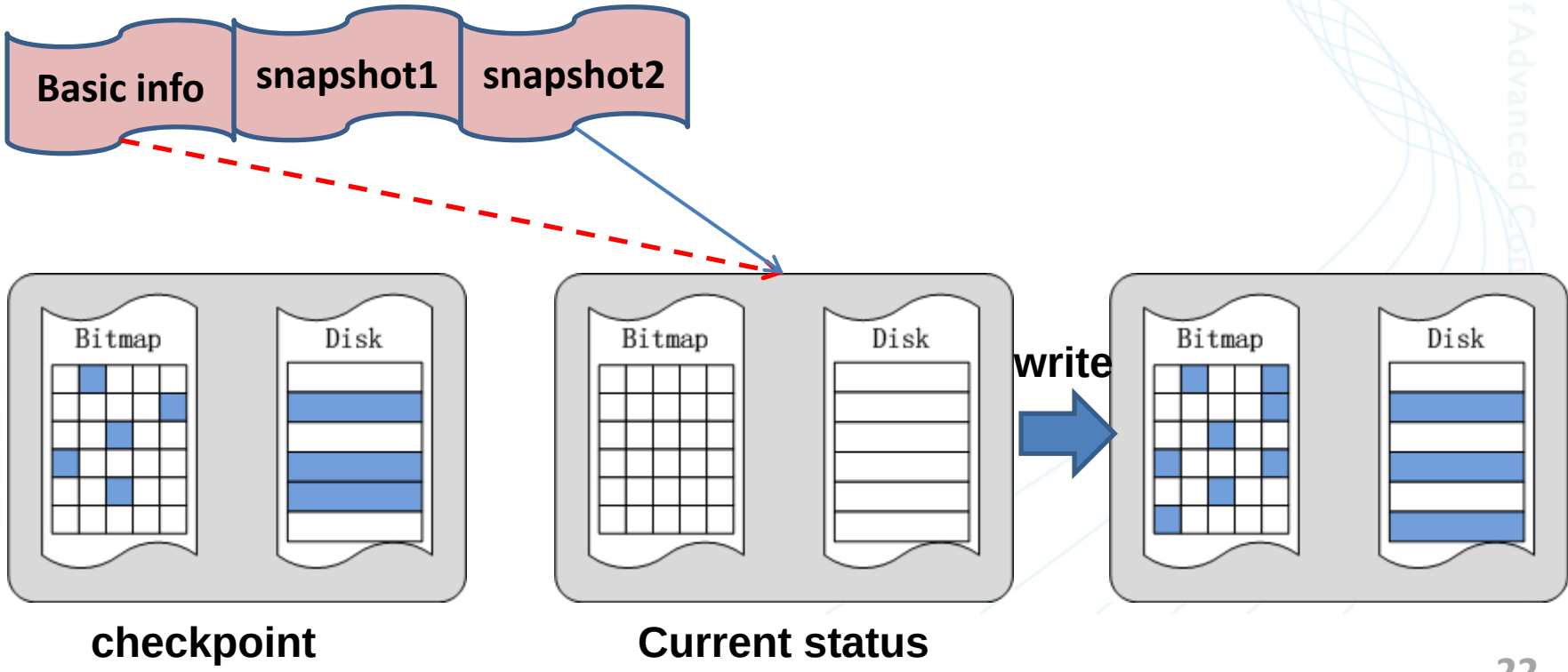
HotSnap skills – disk snapshot

- **ROW based disk snapshot**
  - Only create one bitmap and one null disk image
  - Lightweight metadata based on bitmap
  - Redirect on write



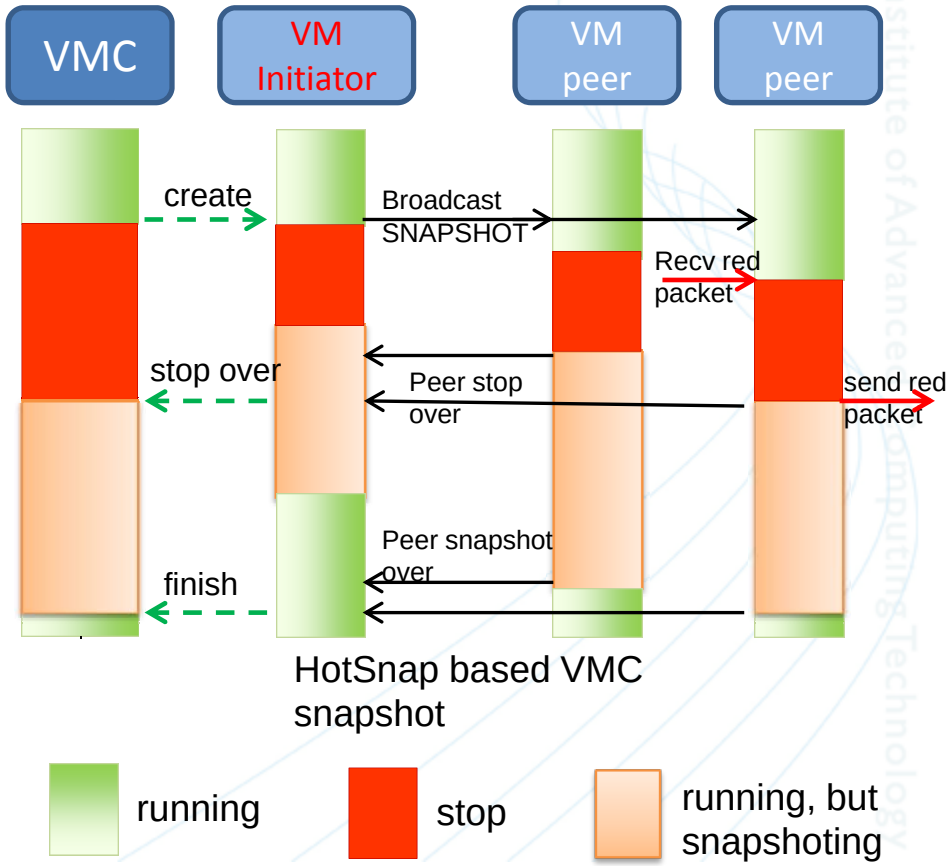
HotSnap skills – disk snapshot

- **ROW based disk snapshot**
  - Only create one bitmap and one null disk image
  - Lightweight metadata based on bitmap
  - Redirect on write



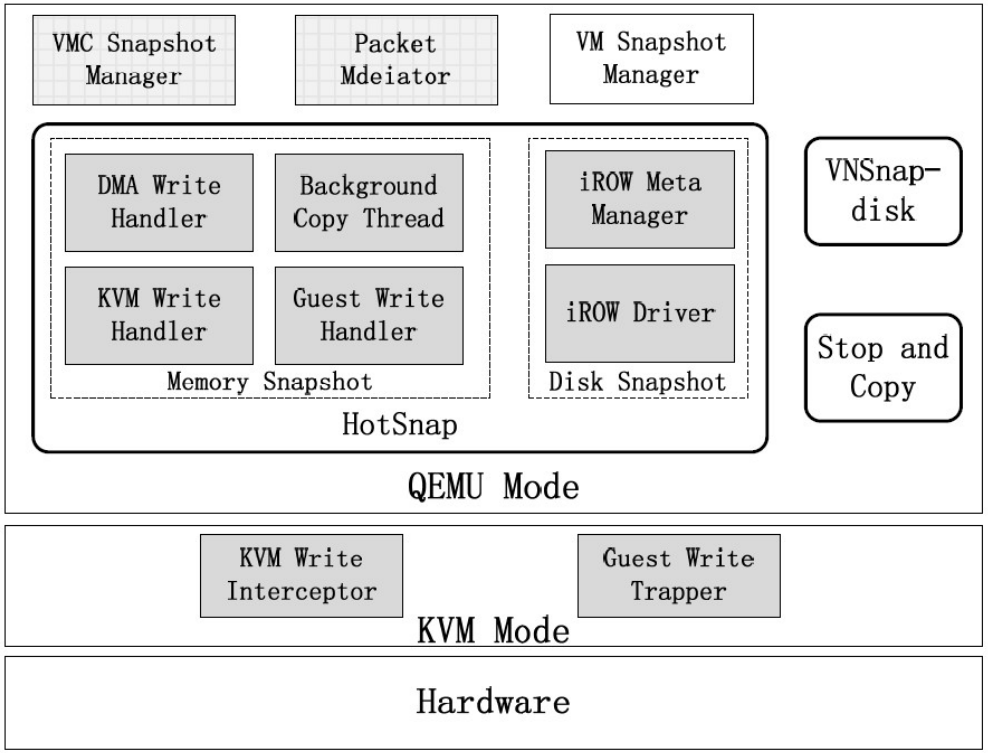
HotSnap skills - Consistency protocol

- Coloring method
  - **Set bit in m_cType in MAC header**
- Protocol
  - **One VM as initiator**
  - **Initiator broadcast SNAPSHOT to VMs**
  - **Peer VM create snapshot, if**
    - **Receive SNAPSHOT**
    - **Receive red packet**
  - **VM colors the packet with red flag after finishing snapshot**
  - **After receiving red packet**
    - **If snapshot is over, continue to run**
    - **Else, create snapshot first, change VM state to red**



HotSnap Architecture

- **Architecture**
  - VM snapshot
    - **Memory Snapshot**
      - **DMA Write Handler**
      - **Background Thread**
      - **Guest Write Handler**
    - **Disk Snapshot**
  - Consistency protocol
    - **VMC Snapshot Manager**
    - **VM Snapshot Manager**
    - **Packet Mediator**



Outline

- Background
- Problems
- Solution & Implementation
- Experimental Results
- Conclusions

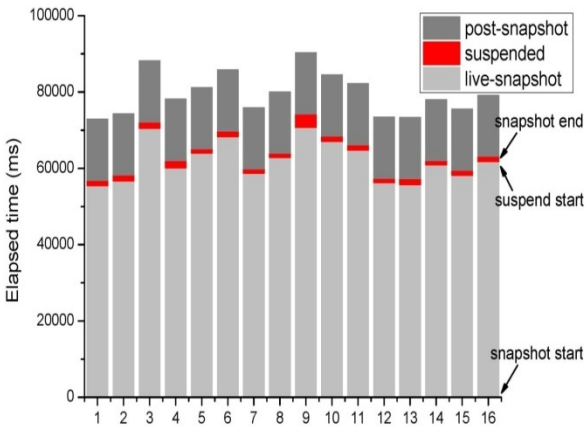
Experimental results

- **VM snapshot (2G memory)**
  - **Duration**: 60s in HotSnap VS. 50s in Pre-copy, related to snapshot size
  - **Downtime**: HotSnap is less than50ms, pre-copy is related to workload
  - **Snapshot size**: HotSnap snapshot size is same to memory size， Pre-copy is much larger.

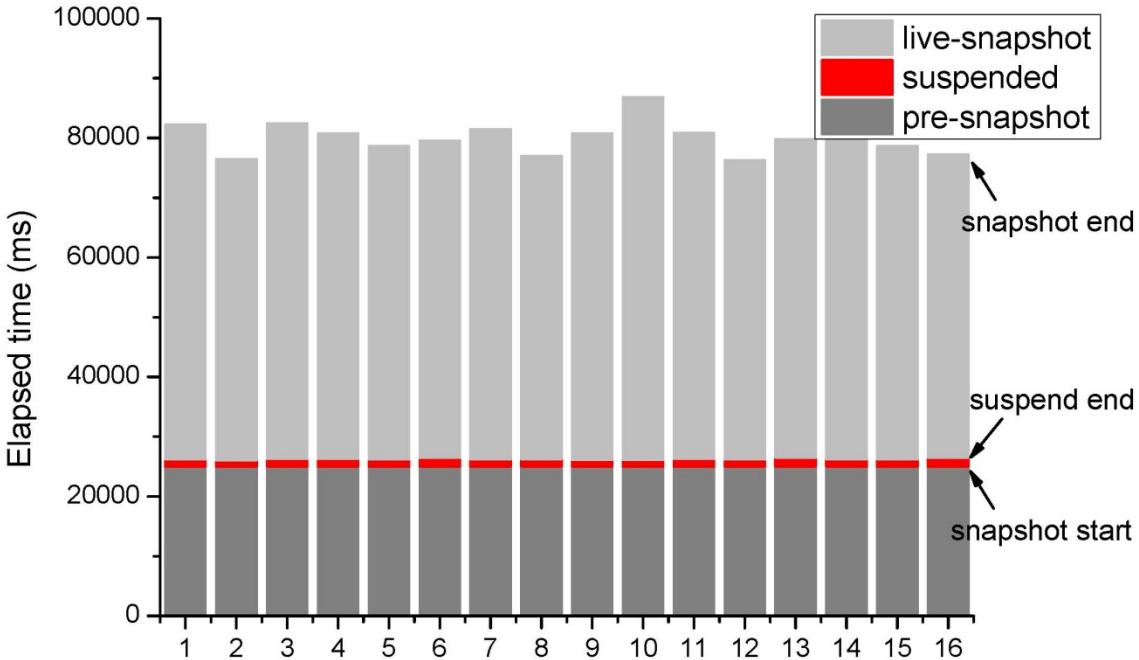
	Duration (s)		Downtime(ms)		Snapshot size(GBytes)	
benchmarks	**Pre-copy**	**Hotsnap**	**Pre-copy**	**Hotsnap**	**Pre-copy**	**Hotsnap**
Idle	51.66	**51.57**	36.83	**31.88**	2.04	**2.02**
Compilzation	61.11	**51.96**	381.72	**34.16**	2.38	**2.02**
Matrix multiplication	51.75	**52.31**	55.73	**35.93**	2.19	**2.02**
Memcached	69.43	**54.72**	150.85	**33.80**	2.41	**2.02**
Dbench	60.76	**50.18**	79.36	**39.36**	2.17	**2.02**

Experimental results

- **VMC snapshot results**
  - 16 2G VMs, 4 physical servers, Distcc to compile kernel
  - Start and end almost at the same time



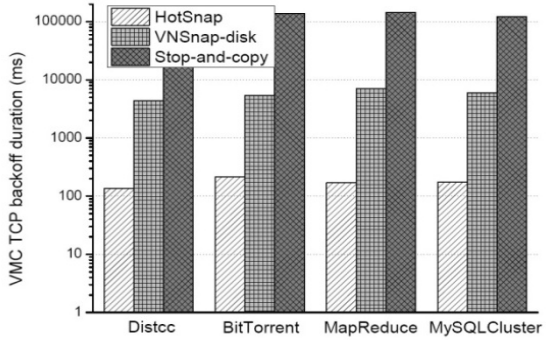
Pre-copy based VMC snapshot



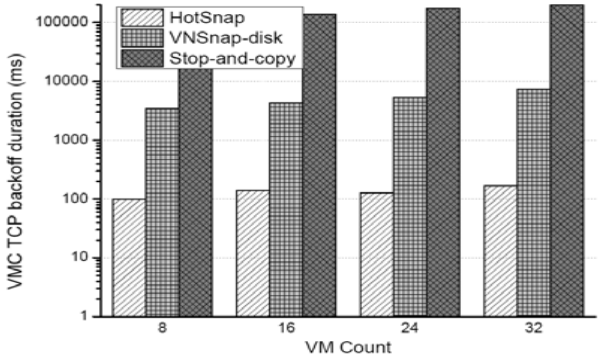
HotSnap for VMC

Experimental results

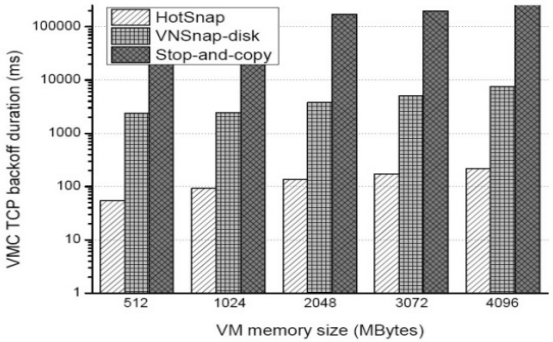
- **TCP backoff**
  - Average of difference of snapshot completion times between each two VMs
  - 16 2G memory VMs, 4 physical servers



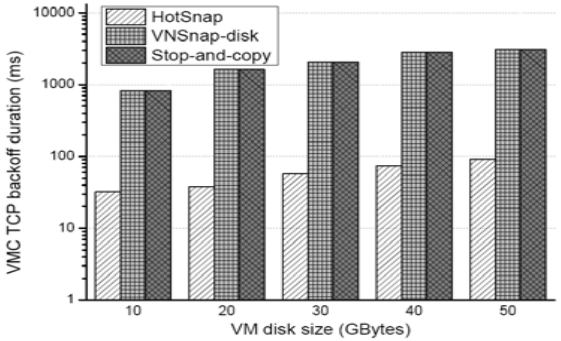
Different workloads



Various VMC size



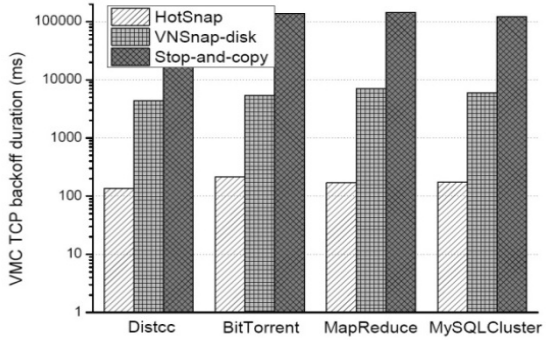
Various memory size



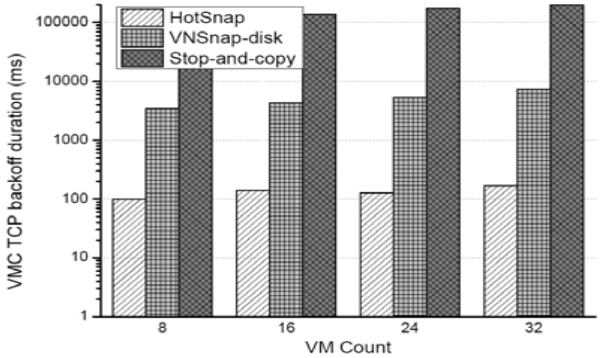
Various disk size

Experimental results

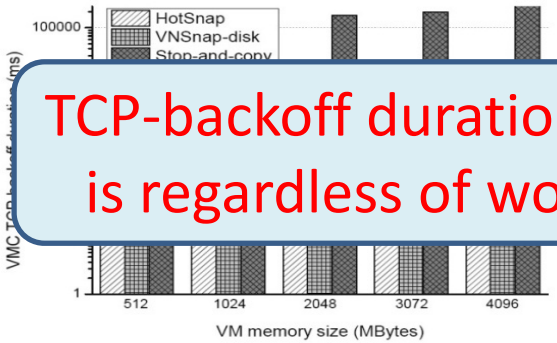
- **TCP backoff**
  - Average of difference of snapshot completion times between each two VMs
  - 16 2G memory VMs, 4 physical servers



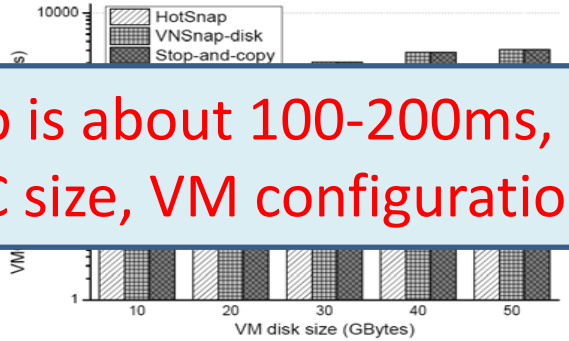
Different workloads



Various VMC size



Various memory size

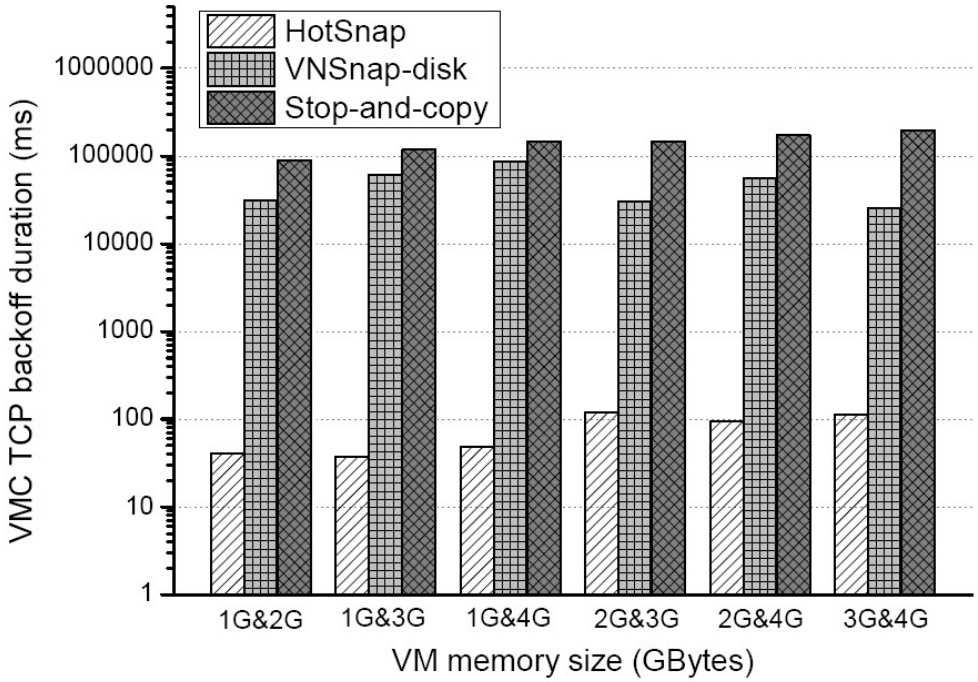


Various disk size

TCP-backoff duration in HotSnap is about 100-200ms, and is regardless of workload, VMC size, VM configurations

Experimental results

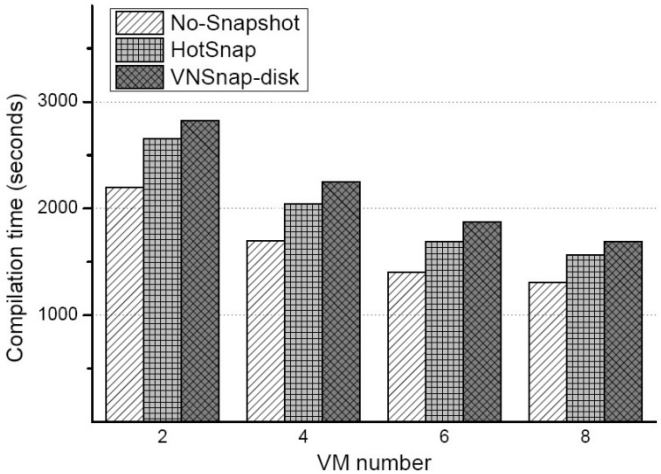
- **Two VMs with different memory size**
  - 1G & 4G has the largest TCP-backoff, 100s
  - Larger difference implies longer TCP-backoff duration for Pre-copy method. HotSnap is regardless of difference.



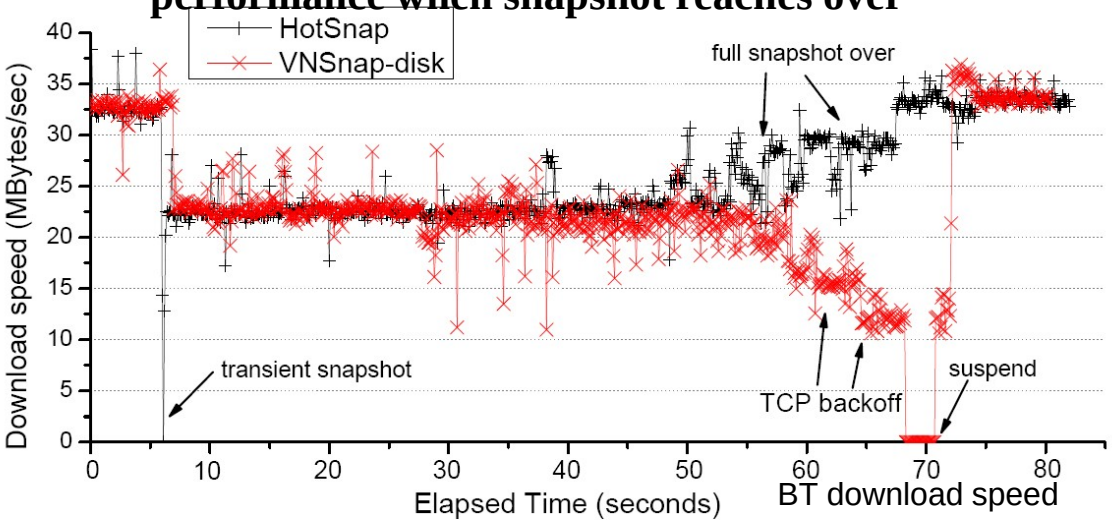
Memory impact on TCP-backoff

Experimental results

- **Performance impact**
  - Kernel compilization
    - **Compared to pre-copy, HotSnap reduces 7%-10% time.**
  - BitTorrent application
    - **16 VMs, one VM as client, others as seeds**
    - **Compared to normal execution, download spe reduce 28%**
    - **Compared to pre-copy, HotSnap shows better performance when snapshot reaches over**



Kernel complication



BT download speed

Outline

- Background
- Problems
- Solution & Implementation
- Experimental Results
- Conclusions

Conclusions

- **HotSnap**
  - Single VM snapshot
    - **Minimize the downtime**
    - **Minimize the difference of snapshot completion times.**
  - Consistency protocol suitable to virtualized environments
- **Experimental results**
  - Single VM snapshot downtime < 100ms
  - 32 VMs, TCP-backoff duration < 1s
  - TCP-backoff is regardless of workload, VMC size, VM configurations
- **Future work**
  - Evaluate HotSnap in real-world applications
  - Reduce the saved amount of VNC snapshot file further

Q&A

ACT lab, Beihang University
{cuilei, libo, zhangyy, lijx}@act.buaa.edu.cn