dist-gem5: Distributed Simulation of Compute Clusters

Mohammad Alian, Umur Darbaz, Gabor Dozsa, Stephan Diestelhorst, Daehoon Kim, Nam Sung Kim
University of Illinois Urbana-Champaign
ARM Ltd., Cambridge, UK
Outline

• motivation
  ✔ accelerating large-scale simulation

• dist-gem5 architecture
  ✔ packet forwarding
  ✔ synchronization
  ✔ checkpointing
  ✔ network model

• evaluation
  ✔ validation, speedup, synchronization overhead

• conclusion
Outline

• motivation
  ✓ accelerating large-scale simulation

• dist-gem5 architecture
  ✓ packet forwarding
  ✓ synchronization
  ✓ checkpointing
  ✓ network model

• evaluation
  ✓ validation, speedup, synchronization overhead

• conclusion
What is gem5 – overview

- full-system, cycle-level, event-driven simulator
- used/maintained at universities and industry

CPU Models
- ARMv7a
- ARMv8
- Atomic
- Timing
- Out of Order
- In Order

ARM ISA Support
- GICv2
- L1-L3
- SCU
- ArchTimer
- PMU

Core Integrated IP

GPU models
- NoMali

Interconnect
- Crossbar
- Snoop filter
- Bridges

Memory
- Flash
- DRAM
- HMC

Simulation support
- Traffic Gen
- Traffic Monitor
- Stream Line
- KVMv7
- FracFact
- Sim Points
- KVMv8
- Power Model Int.
- PCA

IO components
- UHDLCD
- UART
- DMA
- RTC
- UFS
- NVMe
- Timers
- 10Gb NIC

Additional notes:
- ARMv7a
- ARMv8
- GICv2
- L1-L3
- SCU
- ArchTimer
- PMU
- Traffic Gen
- Traffic Monitor
- Stream Line
- KVMv7
- FracFact
- Sim Points
- KVMv8
- Power Model Int.
- PCA
- UHDLCD
- UART
- DMA
- RTC
- UFS
- NVMe
- Timers
- 10Gb NIC
Why dist-gem5?

• performance and power dissipation of a distributed system
  ✔ complex interplay among system components at scale

• need a **full-system, cycle-level** simulator which is **fast** enough to simulate a large-scale computer system

• distributed simulation:
  ✔ simulate a distributed system w/ many simulation hosts
dist-gem5 architecture – high level view

• gem5 processes modeling full systems run in parallel on a cluster of physical machines

• simulated network switch
  ✓ forward packets among the simulated systems
  ✓ synchronize the distributed simulation
  ✓ simulate network topology
Outline

• motivation
  ✓ accelerating large-scale simulation

• dist-gem5 architecture
  ✓ packet forwarding
  ✓ synchronization
  ✓ checkpointing
  ✓ network model

• evaluation
  ✓ validation, speedup, synchronization overhead

• conclusion
dist-gem5 architecture – core components

- Packet forwarding
- Simulated network
- Distributed check-pointing

Accuracy

Correctness

Synchronization
dist-gem5 architecture – core components

- Packet forwarding
- Simulated network
- Distributed check-pointing

Accuracy
Correctness
Synchronization

what is gem5
dist-gem5 architecture
evaluation
conclusion
dist-gem5 architecture – packet forwarding
dist-gem5 architecture – packet forwarding

- Physical host #1
- Physical host #2
- Physical host #3
- Physical switch

- Simulated NIC in gem5 #1
- Simulated system #1
- NIC#1
- Port1

- Simulated NIC in gem5 #2
- Simulated system #2
- NIC#2
- Port2

- Physical NIC in gem5 #3
- Port0
- Port1

- Dist-gem5 architecture – packet forwarding

- What is gem5
- Dist-gem5 architecture
- Evaluation
- Conclusion
dist-gem5 architecture – packet forwarding

simulated packets are embedded into host TCP/IP packets

what is gem5

dist-gem5 architecture

evaluation

conclusion
Asynchronous processing of incoming messages

• simulation thread (main thread)
  ✓ process/insert events in the event queue
  ✓ in case of send pkt event, encapsulate the simulated Ethernet packet in a message and send it out

• receiver thread
  ✓ create for each gem5 process
  ✓ waits for incoming packets
  ✓ creates a recv pkt event and insert it to the event queue
dist-gem5 architecture – core components

- Packet forwarding
- Simulated network
- Distributed check-pointing

Accuracy
Correctness

Synchronization
Need for synchronization

- receiver gem5 can run ahead of sender gem5
  - physical host mismatch
  - different events to be processed
- slowed down receiver gem5 to ensure simulation accuracy
- quantum-based synchronization
Accurate packet forwarding

- **quantum**: interval for periodic synchronization in simulated time
- sync-event flushes inter gem5 communication channels
- if quantum ≤ simulated link delay:
  - expected delivery tick falls inside the next quantum
- optimal quantum size for accurate forwarding == **simulated link delay**
dist-gem5 architecture – core components

packet forwarding

simulated network

distributed check-pointing

synchronization

accuracy

correctness

what is gem5

dist-gem5 architecture

evaluation

conclusion
dist-gem5 architecture – network modeling

simulate in one gem5 process

aggregate switch

top of rack switch #0
  Server #0
  Server #1
  server #2
  server #3
  server #4
  server #5
  server #6
  server #7

top of rack switch #1
  server #8
  server #9
  server #10
  server #11
  server #12
  server #13
  server #14
  server #15

top of rack switch #7
  server #56
  server #57
  server #58
  server #59
  server #60
  server #61
  server #62
  server #63

what is gem5
dist-gem5 architecture
evaluation
conclusion
**Configurable network model**

- configurable baseline Ethernet switch model
  - port number, delay, bandwidth, buffer size

---

**Diagram:**
- SIMULATED PORT
- SIMULATED ETHERLINK
- SIMULATED ETHERSWITCH
- GEM5
- WHAT IS GEM5
- DIST-GEM5 ARCHITECTURE
- EVALUATION
- CONCLUSION

**Abbreviations:**
- MAC
- IN-ORDERQ
- IPORT
- OPORT
- OPART
- IPORT
- OPORT
- P0
- P1
- P7
- P8
- MAC Table

**Legend:**
- Physical host
- Aggregate switch
- Simulated port
- Simulated etherLink
- Simulated etherSwitch

---

**Diagram Details:**
- Top of rack switch #0
- Top of rack switch #1
- Top of rack switch #7
- DistEtherLink
- Physical host

---

**Text:**
- Configurable baseline Ethernet switch model
- Port number, delay, bandwidth, buffer size

---

**Conclusion:**
- What is gem5
- Dist-gem5 architecture
- Evaluation
Outline

• motivation
  ✓ accelerating large-scale simulation

• dist-gem5 architecture
  ✓ packet forwarding
  ✓ synchronization
  ✓ checkpointing
  ✓ network model

• evaluation
  ✓ validation, speedup, synchronization overhead

• conclusion
Methodology – simulation techniques

- For example, simulating a cluster w/ 7 nodes and 1 network switch:

  **single-threaded-gem5**
  - quad core physical host
  - system#6
  - system#4
  - system#2
  - system#0
  - gem5#0

  **parallel-gem5**
  - quad core physical host
  - system#6
  - system#4
  - system#2
  - system#0
  - gem5#6
  - gem5#4
  - gem5#2
  - gem5#0

  **dist-gem5**
  - quad core physical host
  - system#6
  - system#4
  - system#2
  - system#0
  - gem5#6
  - gem5#4
  - gem5#2
  - gem5#0
  - switch
  - system#5
  - system#3
  - system#1
  - gem5#7
  - gem5#5
  - gem5#3
  - gem5#1

what is gem5
dist-gem5 architecture
evaluation
conclusion
Methodology – experimental setup

• focus on off-chip network performance using network intensive applications
  ✓ iperf, memcached, httperf, tcptest, netperf, NAS parallel benchmark

• verification/validation against:
  ✓ single-threaded-gem5
  ✓ physical cluster
    o 4 node cluster w/ AMD A10-5800K

• speedup comparison against:
  ✓ single-threaded-gem5
  ✓ parallel-gem5

<table>
<thead>
<tr>
<th>category</th>
<th>gem5 configuration</th>
</tr>
</thead>
<tbody>
<tr>
<td>O3 core</td>
<td>4 cores; 4 way superscalar</td>
</tr>
<tr>
<td>memory</td>
<td>8GB DDR3 1600 MHz</td>
</tr>
<tr>
<td>network</td>
<td>Intel GbE NIC; 1 μs Link latency</td>
</tr>
<tr>
<td>OS</td>
<td>Linux Ubuntu 14.04 (Kernel 4.3)</td>
</tr>
</tbody>
</table>
Verification

• same node/network config
  ✓ dist-gem5 generates identical simulation statistics compared to single-threaded-gem5
  ✓ different cluster sizes
Validation – network latency and bandwidth

- iperf (left) and memcached (right)
- follows the behavior of physical setup
- 17.5% lower response time for memcached
Speedup – simulation time reduction

• running httpperf on each simulated node sending fixed number of requests to a unique simulated node (apache server)
• compared with single-threaded-gem5
• dist-gem5 simulating 63 nodes on 16 physical hosts is
  ✓ $83.1 \times$ faster than single-threaded-gem5
  ✓ $12.8 \times$ faster than parallel-gem5

speedup of parallel-gem5 saturates!
Scalability – simulation time vs. simulated cluster size

- simulation time increase for simulating 64 vs. 3 nodes:
  - 57.3× for Single-threaded-gem5
  - 23.9× for parallel-gem5
  - 1.9× for dist-gem5

**dist-gem5 scales well!**
Synchronization overhead

- sweep synchronization quantum size
- # of http req remains near constants
  - maximum 2.6% variance
  - almost the same amount of work done at each quantum size
- simulation time improvement
  - 4.9% from 0.5 µs to 1 µs
  - 15.7% from 0.5 µs to 128 µs

dist-gem5 synchronization is efficient!
Conclusion

• dist-gem5 is a distributed version of gem5 for modeling computer clusters
  ✓ validated against a physical cluster
  ✓ accurate/deterministic
  ✓ rich off-chip network modeling
  ✓ 83.1x speedup over single-threaded-gem5 simulating a 63 node cluster

• integrated to mainstream gem5
  ✓ available at gem5.org
  ✓ enabled via “--dist” command line option

• developed/maintained by university and industry
Thank You