BabelFish: Fusing Address Translations for Containers

Dimitrios Skarlatos, Umur Darbaz, Bhargava Gopireddy, Nam Sung Kim, and Josep Torrellas
University of Illinois at Urbana-Champaign
skarlat2@illinois.edu

ISCA 2020
Conventional Cloud Computing

Virtual Machines

- App
- Lib
- Guest OS
- Hypervisor
- Hardware

Hardware

Hypervisor

Guest OS

Lib

App
New Era in Cloud Computing

1. Lightweight
2. Faster bringup
3. Higher consolidation

Containers

OS automatically shares resources!
Containers in the Cloud

- Replicated Containers
- Scalability
- Fault Tolerance
- Load Balancing

High number of containers
Very frequent context switches
Serverless & Function-as-a-Service (FaaS)
  - Cold start effects
  - Very short runtime

Pod
- Application Binary
- Library
- Container Engine
- Operating System
- Hardware
- Shared Data Pages

kubernetes
Problem: Replicated VA → PA Translations

1. TLB thrashing
2. Redundant page table work

- Scalability
- Fault Tolerance
- Load Balancing

Replicated Containers

High number of containers
- Very frequent context switches
- Serverless & Function-as-a-Service (FaaS)
  - Cold start effects
  - Very short runtime

Replicated VA → PA Translations

Pod

Application Binary

Library

Shared Data Pages

Hardware

Operating System

Application Binary

Kubernetes
Solution: Share VA→PA Translations

1. Minimize TLB thrashing
2. Eliminate redundant page table work

Replicated Containers
Scalability
Fault Tolerance
Load Balancing

Shared Data Pages
Library
Application Binary
Pod

Function-as-a-Service (FaaS)
- Cold start effects
- Very short runtime
Contribution: BabelFish

HW and OS support to share translations across containers
1. Introduce Container Context IDentifiers (CCID)
2. Extend TLB design
3. Share page tables

Performance improvement
Data-serving: 11%-18%
Compute: 11%
Function-as-a-Service: 10-55%
Container bring-up: 8%

“The Babel fish is small, yellow, leech-like, and probably the oddest thing in the Universe. The practical upshot of all this is that if you stick a Babel fish in your ear you can instantly understand anything said to you in any form of language.”

The Hitchhiker’s Guide to the Galaxy – Douglas Adams
Problem: TLB Bloat
Problem: TLB Bloat

Main Memory

Container A

Core

L1 Cache

L2 Cache

L3 Cache

Page Tables A

VA1

PA4

TLB

Issue LD VA 1

TLB Miss → “Page Walk” = Fetch entry from page table

ILLINOIS
Problem: TLB Bloat

Container A

Core

L1 Cache

L2 Cache

L3 Cache

Main Memory

Page Tables A

PA 4

Existing Process Context Identifiers (PCID)
Problem: TLB Bloat

What if we could identify shared entries?

TLB Miss → “Page Walk” = Fetch entry from page table
Problem: TLB Bloat

What if we could identify shared entries?
Solution: BabelFish TLB

Unlimited sharing of translations among any number of containers!
Solution: BabelFish TLB

Container Context Identifier (CCID)

Container A  Container B  Container C

Core  L1 Cache  L2 Cache  L3 Cache

TLB

Main Memory

Page Tables B

PrivateCopy (PC)
VA→PA shared for some containers & private for others
Solution: BabelFish TLB

Container Context Identifier (CCID)

Container A  Container B  Container C

Core  L1 Cache  L2 Cache  L3 Cache

TLB

PrivateCopy (PC)
VA→PA shared for some containers & private for others
Problem: Page Table Entry Bloat
Problem: Page Table Entry Bloat
Problem: Lazy Page Table Management
Problem: Lazy Page Table Management

Main Memory

Container A

Core

L1 Cache

L2 Cache

L3 Cache

TLB

Page Tables A

Page Tables B

VA 1
PA 4
P

Present bit currently cleared
Problem: Lazy Page Table Management

Container A

Core

L1 Cache

L2 Cache

L3 Cache

TLB

Main Memory

Page Tables A

Page Tables B

Storage

Issue LD VA 1

Page Fault!
Problem: Lazy Page Table Management

- **Core**
- **L1 Cache**
- **L2 Cache**
- **L3 Cache**
- **TLB**
- **Main Memory**
- **Storage**
- **Page Tables A**
  - VA 1
  - PA 4
  - Present bit set
- **Page Tables B**
  - VA 1
  - PA 4
  - Present bit remains cleared

**Container A**

- Issue LD VA 1
- Page Fault!
Problem: Lazy Page Table Management

Container B

Core

L1 Cache

L2 Cache

L3 Cache

TLB

Main Memory

Page Tables A

Page Tables B

Present bit set

Issue LD VA 1

Page Fault!
Conventional Page Tables

Virtual Address:

- 47 … 39: 9-bits
- 38 … 30: 9-bits
- 29 … 21: 9-bits
- 20 … 12: 9-bits
- 11 … 0: 12-bits

Flow:

- CR3 → PGD
- PGD → PUD
- PUD → PMD
- PMD → PTE
- PTE → To TLB

Container A
Conventional Page Tables

Container A

Container B

**Conventional Page Tables**

Virtual Address:

<table>
<thead>
<tr>
<th>47 … 39</th>
<th>38 … 30</th>
<th>29 … 21</th>
<th>20 … 12</th>
<th>11 … 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>9-bits</td>
<td>9-bits</td>
<td>9-bits</td>
<td>9-bits</td>
<td>12-bits</td>
</tr>
</tbody>
</table>

- **PGD** (Page Directory) for Container A
- **PUD** (Page Unit Directory) for Container A
- **PMD** (Page Macro Directory) for Container A
- **pte_t** (Page Table Entry) for Container A
- **CR3** (Control Register)

- **PGD** (Page Directory) for Container B
- **PUD** (Page Unit Directory) for Container B
- **PMD** (Page Macro Directory) for Container B
- **pte_t** (Page Table Entry) for Container B
- **CR3** (Control Register)
Conventional Page Tables

Container A

Container B

Virtual Address

47 ... 39  38 ... 30  29 ... 21  20 ... 12  11 ... 0

9-bits  9-bits  9-bits  9-bits  12-bits

To TLB

To TLB

pgd

pte_t

pte_t
Solution: BabelFish Page Table Sharing

Container A

Container B

Virtual Address

47 … 39  38 … 30  29 … 21  20 … 12  11 … 0

9-bits  9-bits  9-bits  9-bits  12-bits

Virtual Address

CR3₀

PGD

PUD

PMD

CR3₁

PGD

PUD

PMD

PTE

pte_t

To TLB
Example

Container A: Runs on Core 0, requests VA → PA
Container B: Runs on Core 1, requests VA → PA
Container C: Runs on Core 0, requests VA → PA
Example

Conventional

Issue LD V1
Core 0

L1 TLB Miss
L2 TLB Miss
PWC Miss

PGD Walk
PUD Walk
PMD Walk
PTE Walk

Page Walk Cache (PWC)

L2 Cache Miss, L3 Cache Miss, Main Memory Hit

Time

Page Fault
## Example

### Conventional

<table>
<thead>
<tr>
<th>Issue LD V1 Core 0</th>
<th>L1 TLB Miss</th>
<th>L2 TLB Miss</th>
<th>PWC Miss</th>
<th>PGD Walk</th>
<th>PUD Walk</th>
<th>PMD Walk</th>
<th>PTE Walk</th>
<th>Page Fault</th>
</tr>
</thead>
<tbody>
<tr>
<td>Issue LD V1 Core 1</td>
<td>L1 TLB Miss</td>
<td>L2 TLB Miss</td>
<td>PWC Miss</td>
<td>PGD Walk</td>
<td>PUD Walk</td>
<td>PMD Walk</td>
<td>PTE Walk</td>
<td>Page Fault</td>
</tr>
</tbody>
</table>

Time
### Example

| Conventional                                                                 |
|                                                                             |
| ![Image](image1.png)                                                        |
|                                                                             |
| **Issue LD V1**                                                              |
| Core 0                                                                      |
| L1 TLB Miss                                                                 |
| L2 TLB Miss                                                                 |
| PWC Miss                                                                    |
| PGD Walk                                                                    |
| PUD Walk                                                                    |
| PMD Walk                                                                    |
| PTE Walk                                                                    |
| **Page Fault**                                                               |
|                                                                             |
| **Issue LD V1**                                                              |
| Core 1                                                                      |
| L1 TLB Miss                                                                 |
| L2 TLB Miss                                                                 |
| PWC Miss                                                                    |
| PGD Walk                                                                    |
| PUD Walk                                                                    |
| PMD Walk                                                                    |
| PTE Walk                                                                    |
| **Page Fault**                                                               |
|                                                                             |
| **Issue LD V1**                                                              |
| Core 0                                                                      |
| L1 TLB Miss                                                                 |
| L2 TLB Miss                                                                 |
| PWC Miss                                                                    |
| PGD Walk                                                                    |
| PUD Walk                                                                    |
| PMD Walk                                                                    |
| PTE Walk                                                                    |
| **Page Fault**                                                               |
### Example

<table>
<thead>
<tr>
<th>Time</th>
<th>Issue LD V1 Core 0</th>
<th>L1 TLB Miss</th>
<th>L2 TLB Miss</th>
<th>PWC Miss</th>
<th>PGD Walk</th>
<th>PUD Walk</th>
<th>PMD Walk</th>
<th>PTE Walk</th>
<th>Page Fault</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Conventional</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Issue LD V1 Core 1</td>
<td>L1 TLB Miss</td>
<td>L2 TLB Miss</td>
<td>PWC Miss</td>
<td>PGD Walk</td>
<td>PUD Walk</td>
<td>PMD Walk</td>
<td>PTE Walk</td>
<td>Page Fault</td>
</tr>
<tr>
<td></td>
<td>BabelFish</td>
<td>L1 TLB Miss</td>
<td>L2 TLB Miss</td>
<td>PWC Miss</td>
<td>PGD Walk</td>
<td>PUD Walk</td>
<td>PMD Walk</td>
<td>PTE Walk</td>
<td>Page Fault</td>
</tr>
<tr>
<td></td>
<td>Issue LD V1 Core 0</td>
<td>L1 TLB Miss</td>
<td>L2 TLB Miss</td>
<td>PWC Miss</td>
<td>PGD Walk</td>
<td>PUD Walk</td>
<td>PMD Walk</td>
<td>PTE Walk</td>
<td>Page Fault</td>
</tr>
</tbody>
</table>

**ILLINOIS**
### Example

<table>
<thead>
<tr>
<th>Time</th>
<th>Issue LD V1 Core 0</th>
<th>Issue LD V1 Core 1</th>
<th>Issue LD V1 Core 0</th>
<th>Issue LD V1 Core 1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L1 TLB Miss</td>
<td>L1 TLB Miss</td>
<td>L1 TLB Miss</td>
<td>L1 TLB Miss</td>
</tr>
<tr>
<td></td>
<td>L2 TLB Miss</td>
<td>L2 TLB Miss</td>
<td>L2 TLB Miss</td>
<td>L2 TLB Miss</td>
</tr>
<tr>
<td></td>
<td>PWC Miss</td>
<td>PWC Miss</td>
<td>PWC Miss</td>
<td>PWC Miss</td>
</tr>
<tr>
<td></td>
<td>PGD Walk</td>
<td>PGD Walk</td>
<td>PGD Walk</td>
<td>PGD Walk</td>
</tr>
<tr>
<td></td>
<td>PUD Walk</td>
<td>PUD Walk</td>
<td>PUD Walk</td>
<td>PUD Walk</td>
</tr>
<tr>
<td></td>
<td>PMD Walk</td>
<td>PMD Walk</td>
<td>PMD Walk</td>
<td>PMD Walk</td>
</tr>
<tr>
<td></td>
<td>PTE Walk</td>
<td>PTE Walk</td>
<td>PTE Walk</td>
<td>PTE Walk</td>
</tr>
<tr>
<td></td>
<td>Page Fault</td>
<td>Page Fault</td>
<td>Page Fault</td>
<td>Page Fault</td>
</tr>
</tbody>
</table>

#### Conventional

- Core 0: L1 TLB Miss, L2 TLB Miss, PWC Miss, PGD Walk, PUD Walk, PMD Walk, PTE Walk, Page Fault
- Core 1: L1 TLB Miss, L2 TLB Miss, PWC Miss, PGD Walk, PUD Walk, PMD Walk, PTE Walk, Page Fault

#### BabelFish

- Core 0: L1 TLB Miss, L2 TLB Miss, PWC Miss, PGD Walk, PUD Walk, PMD Walk, PTE Walk, Page Fault
- Core 1: L1 TLB Miss, L2 TLB Miss, PWC Miss, PGD Walk, PUD Walk, PMD Walk, PTE Walk, Page Fault

- L2 Cache Miss, L3 Cache Hit
Example

Conventional

<table>
<thead>
<tr>
<th>Issue LD V1</th>
<th>L1 TLB Miss</th>
<th>L2 TLB Miss</th>
<th>PWC Miss</th>
<th>PGD Walk</th>
<th>PUD Walk</th>
<th>PMD Walk</th>
<th>PTE Walk</th>
<th>Page Fault</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Core 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

BabelFish

<table>
<thead>
<tr>
<th>Issue LD V1</th>
<th>L1 TLB Miss</th>
<th>L2 TLB Miss</th>
<th>PWC Miss</th>
<th>PGD Walk</th>
<th>PUD Walk</th>
<th>PMD Walk</th>
<th>PTE Walk</th>
<th>Page Fault</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Core 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

L2 Cache Miss, L3 Cache Hit

ILLINOIS
Example

Conventional

<table>
<thead>
<tr>
<th>Issue LD V1 Core 0</th>
<th>L1 TLB Miss</th>
<th>L2 TLB Miss</th>
<th>PWC Miss</th>
<th>PGD Walk</th>
<th>PUD Walk</th>
<th>PMD Walk</th>
<th>PTE Walk</th>
<th>Page Fault</th>
</tr>
</thead>
</table>

BabelFish

| Core 0 | Miss | Miss | Miss | Walk | Walk | Walk | Walk | Page Fault |

<table>
<thead>
<tr>
<th>Issue LD V1 Core 1</th>
<th>L1 TLB Miss</th>
<th>L2 TLB Miss</th>
<th>PWC Miss</th>
<th>PGD Walk</th>
<th>PUD Walk</th>
<th>PMD Walk</th>
<th>PTE Walk</th>
</tr>
</thead>
</table>

| Issue LD V1 Core 0 | L1 TLB Hit | | | | | | |

Result: Faster Execution Time!

L2 Cache Miss, L3 Cache Hit
Workloads & Methodology

Workloads:

- Data serving: ArangoDB, MongoDB, HTTPd
- Compute: GraphChi, FIO
- FaaS: Parse, Hash, Marshal (both dense and sparse)

Full system simulations with Simics + SST
Page Table Sharing for Two Containers

5-minute run on a real machine
Page Table Sharing for Two Containers

5-minute run on a real machine
Page Table Sharing for Two Containers

5-minute run on a real machine
Page Table Sharing for Two Containers

5-minute run on a real machine
Page Table Sharing for Two Containers

5-minute run on a real machine
Page Table Sharing for Two Containers

Page Sharing is Common → Major Reduction in Translations

5-minute run on a real machine
Performance

- **Data Serving**
  - Mean: 10%
  - 95th Percentile: 15%

- **Compute**
  - Mean: 12%
  - 95th Percentile: 13%

- **Dense Functions**
  - Mean: 11%
  - 95th Percentile: 12%

- **Sparse Functions**
  - Mean: 55%
  - 95th Percentile: 55%

- **Bring-up**
  - Mean: 8%
  - 95th Percentile: 9%
Major Performance Gains
→ Both from TLB and page tables

Performance

- Data Serving
- Compute
- Dense Functions
- Sparse Functions
- Bring-up
More in the Paper

Copy-on-Write pages
Address Space Layout Randomization (ASLR) for security
Other security considerations
More evaluation
  Area, energy, sharing, comparison vs larger TLB
Takeaway: **BabelFish**

Sharing Translations in TLB and Page Tables

Introduces Container Context ID (CCID) at OS and HW
TLB extensions & Page Table sharing support
Substantial speedup in application execution and container bring-up

Future work → CCID in OS and HW
- Virtualized environments and nested page tables
- Other resources
BabelFish: Fusing Address Translations for Containers

Dimitrios Skarlatos, Umur Darbaz, Bhargava Gopireddy, Nam Sung Kim, and Josep Torrellas
University of Illinois at Urbana-Champaign
skarlat2@illinois.edu

ISCA 2020