Distributed Systems in Practice, in Theory

Aysylu Greenberg

March 8th, 2016

HAPPY INTERNATIONAL WOMEN'S DAY 8th MARCH

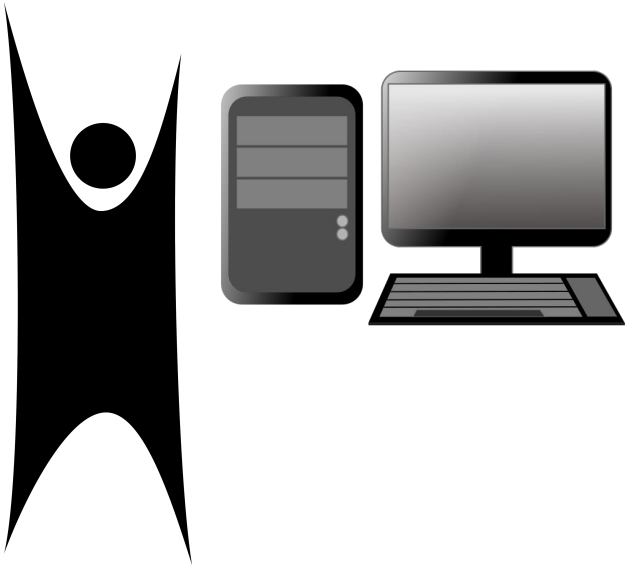
www.jeyjoo.com/gallery

Aysylu Greenberg

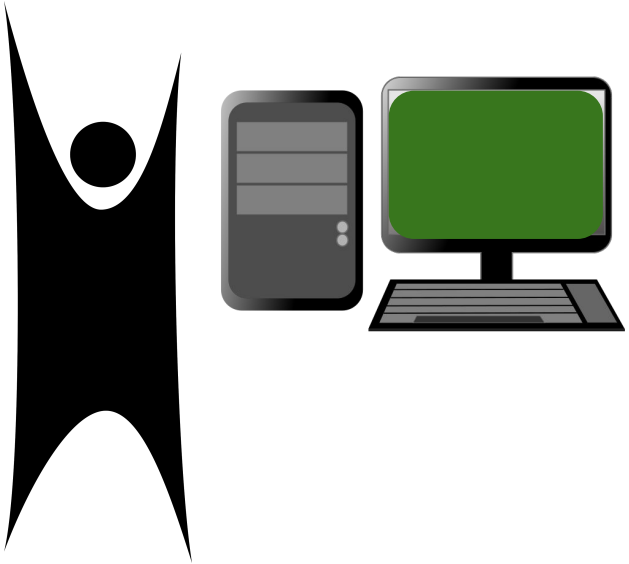
Google

@aysylu22

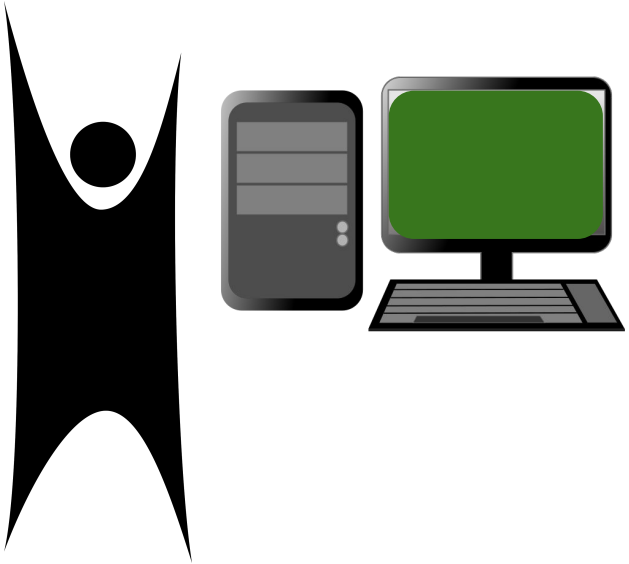
Towards Distributed Build System

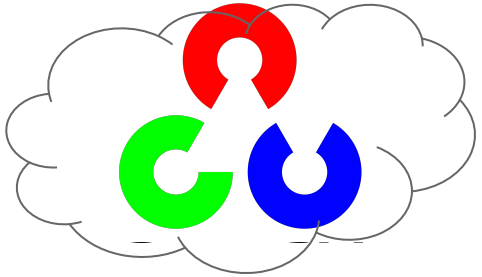


Towards Distributed Build System



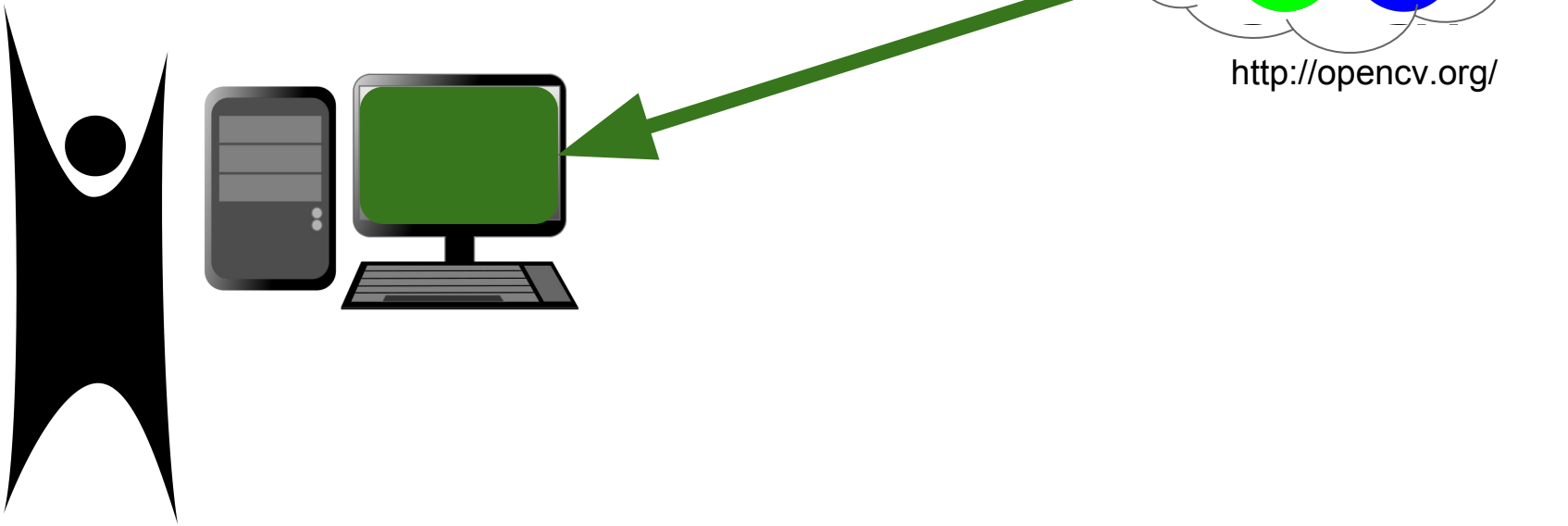
Towards Distributed Build System



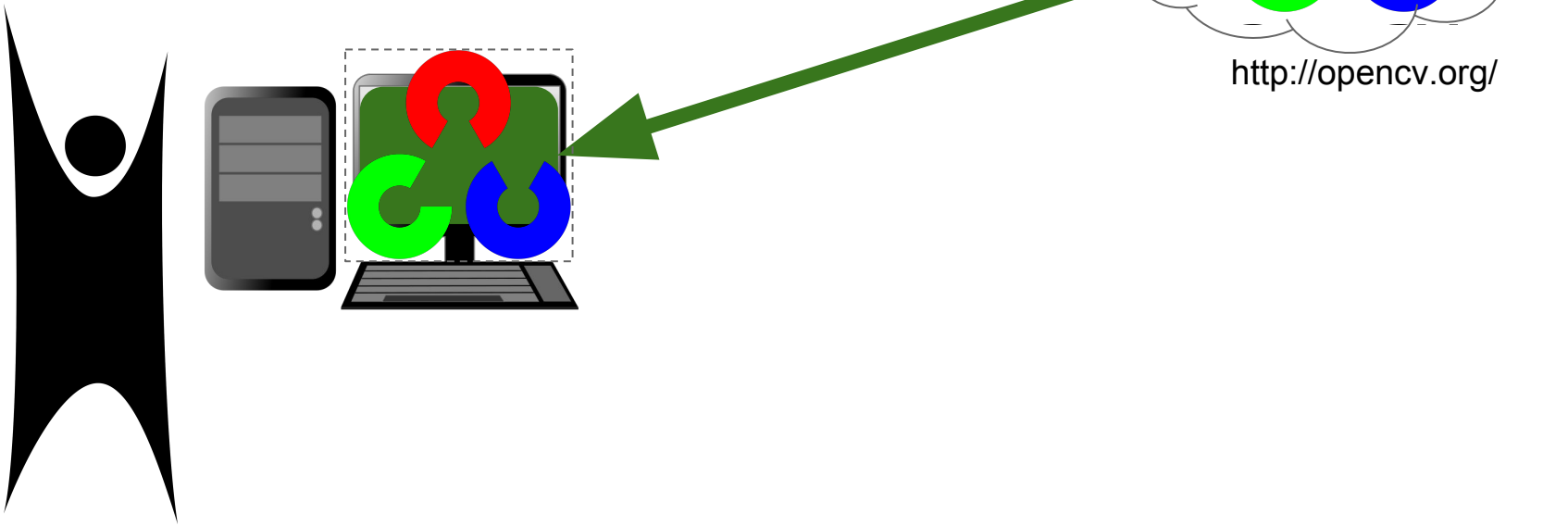


http://opencv.org/

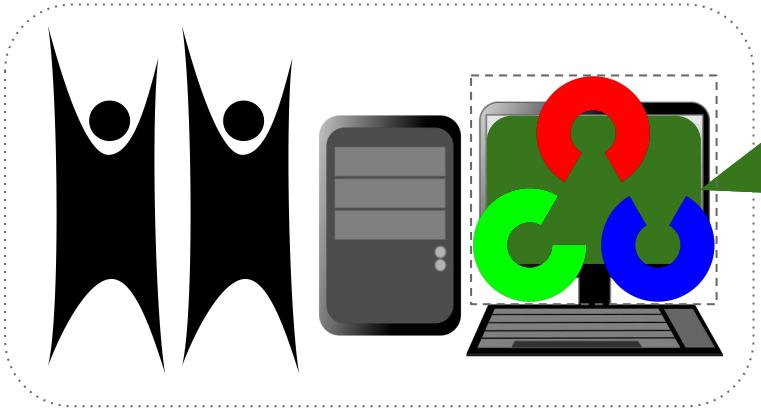
Towards Distributed Build System

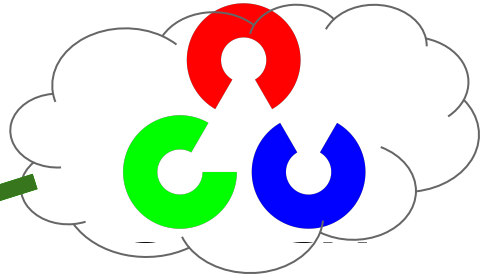


Towards Distributed Build System



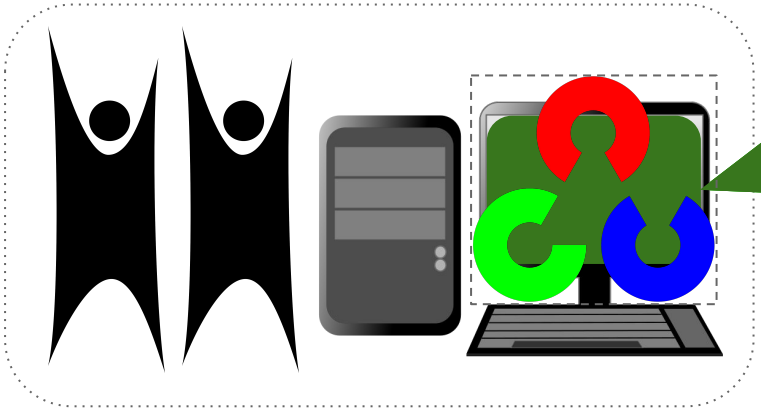
Towards Distributed Build System

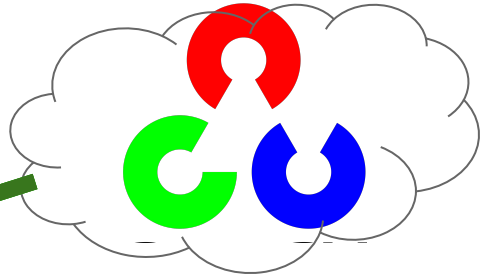




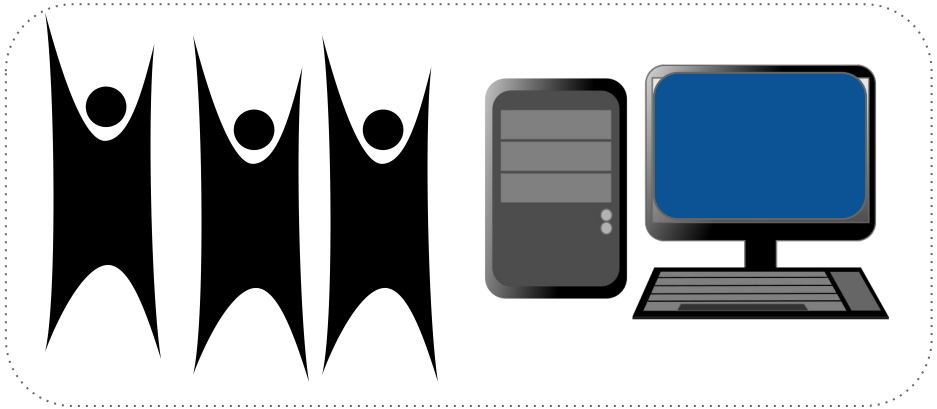
http://opencv.org/

Towards Distributed Build System

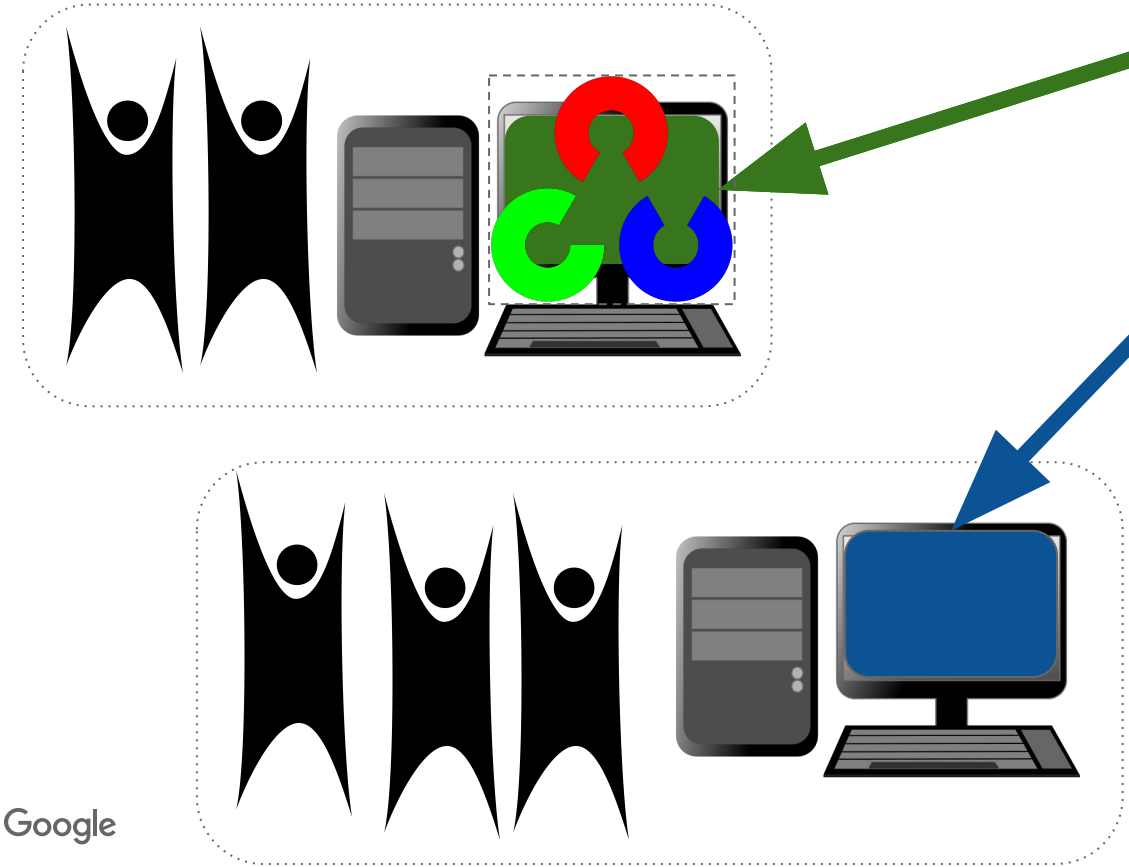




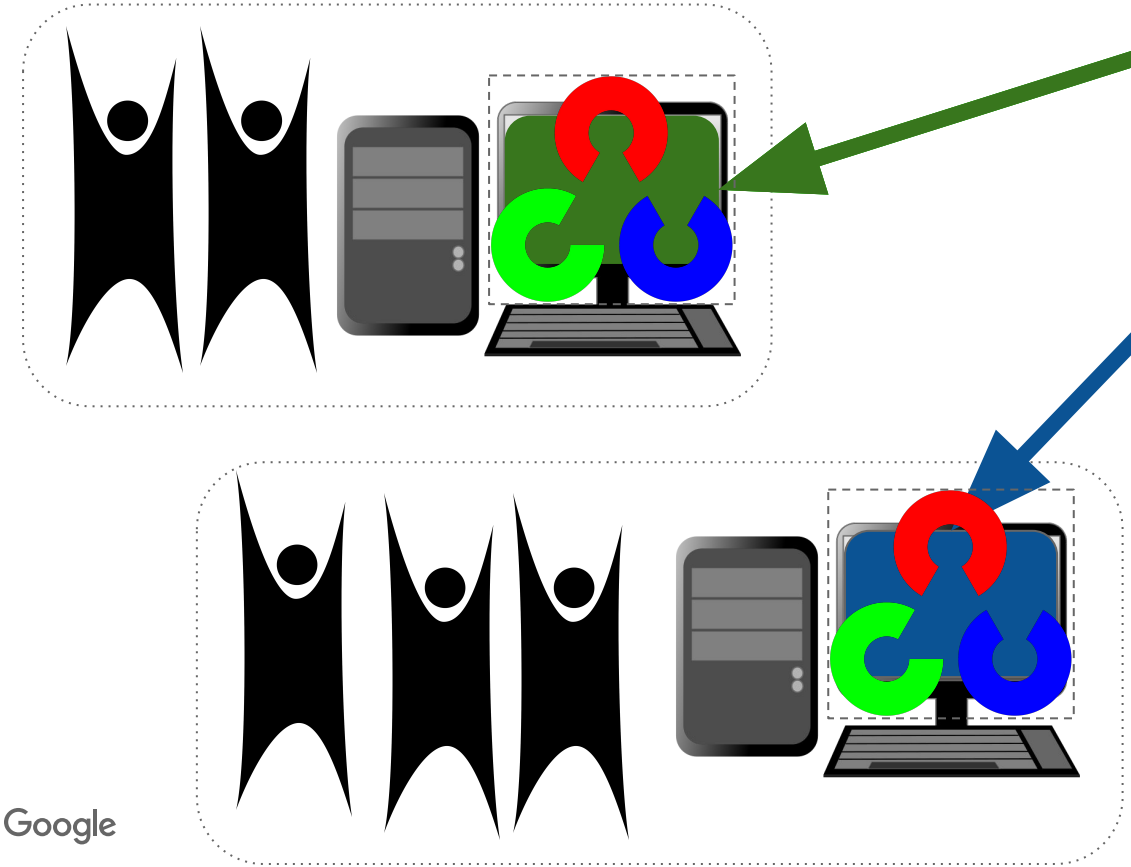
http://opencv.org/



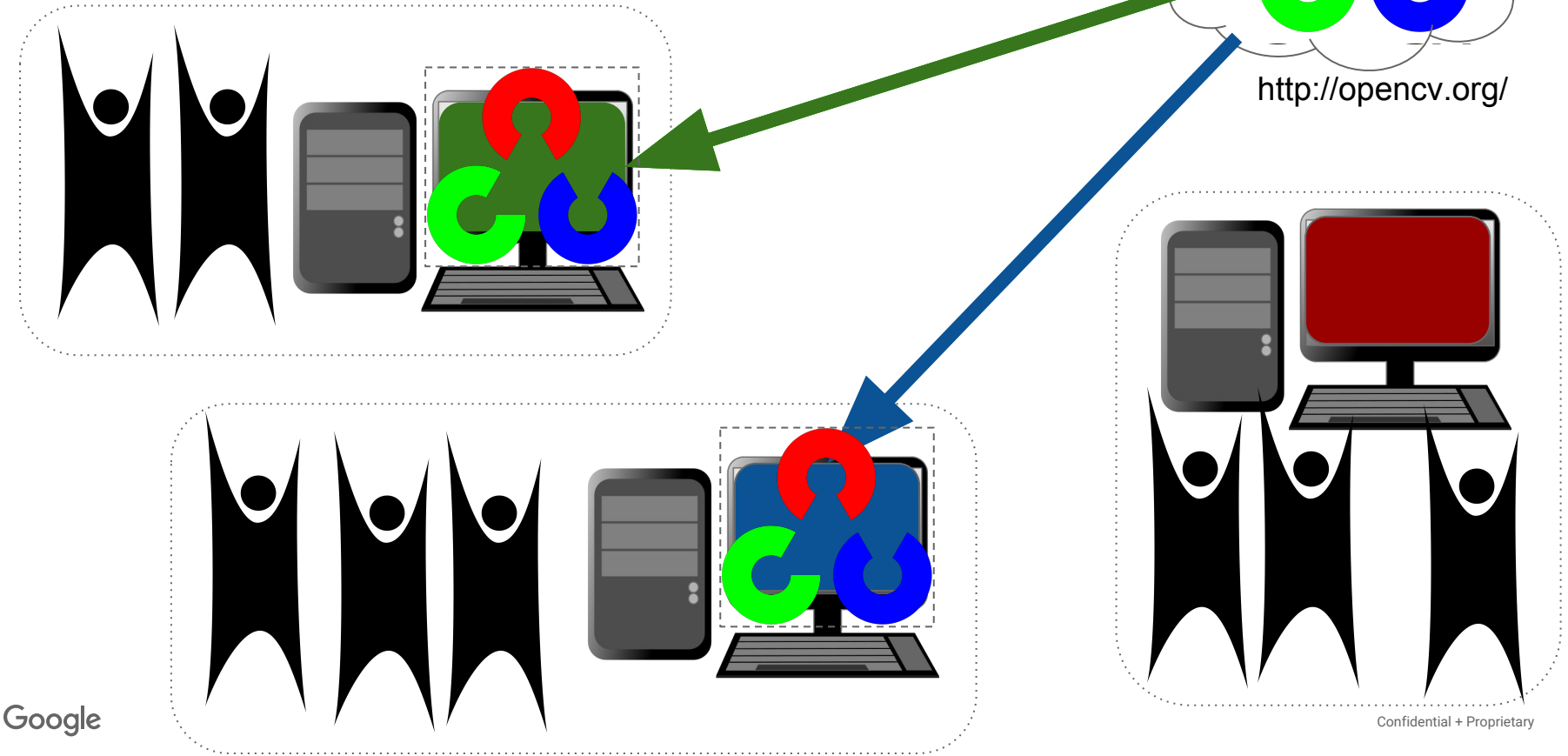
Towards Distributed Build System



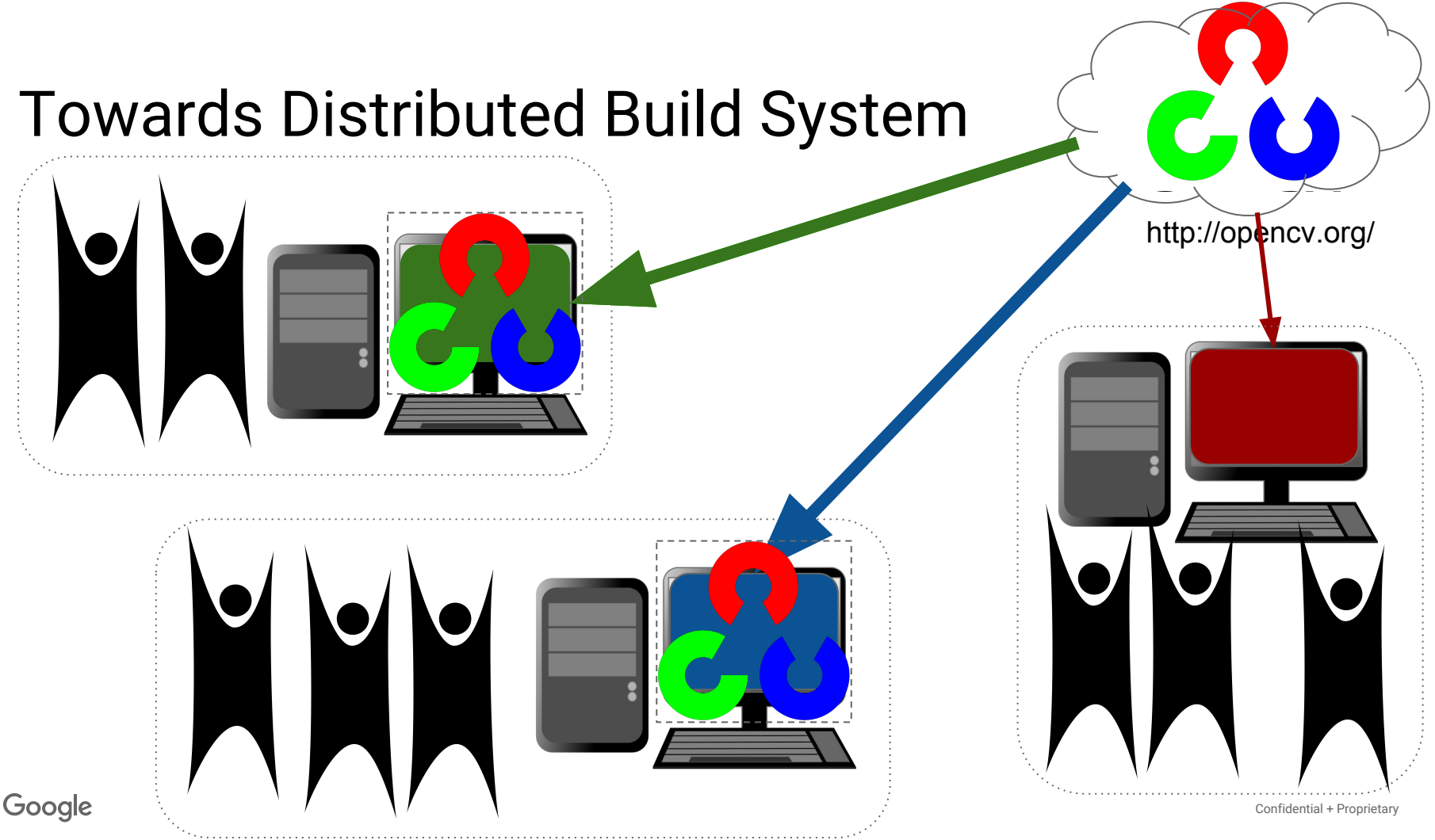
Towards Distributed Build System



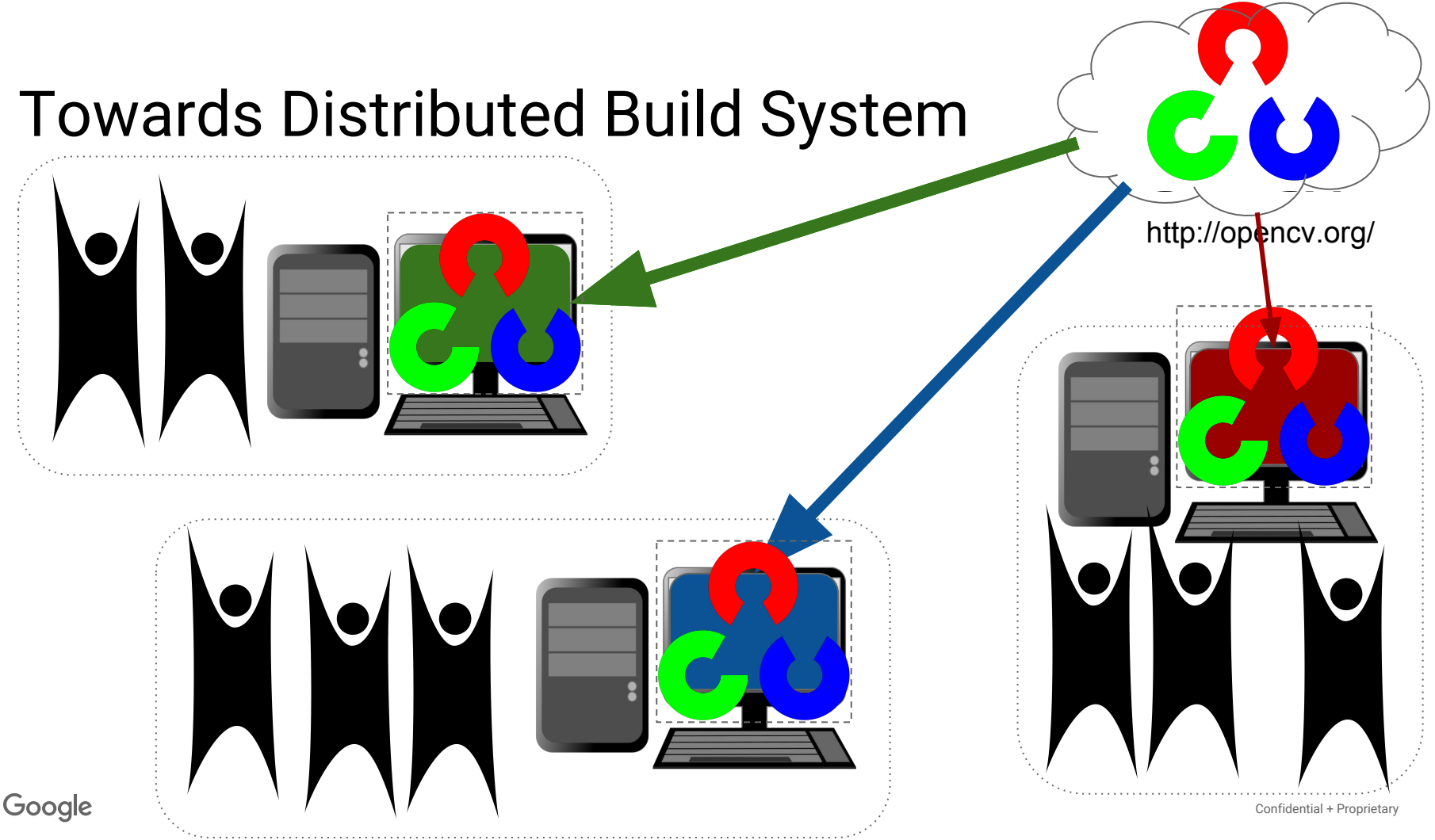
Towards Distributed Build System



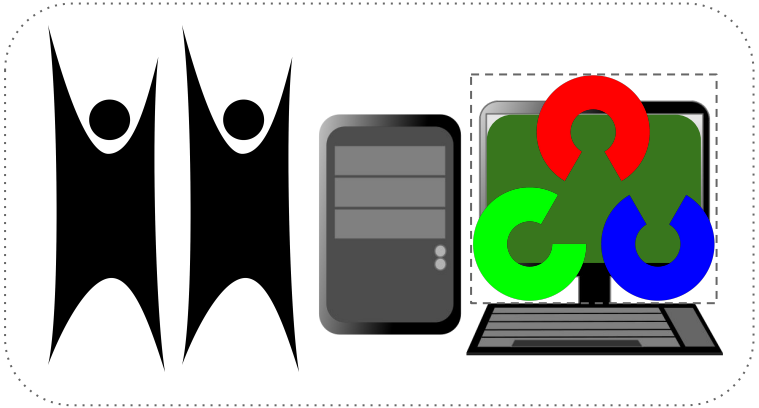
Towards Distributed Build System

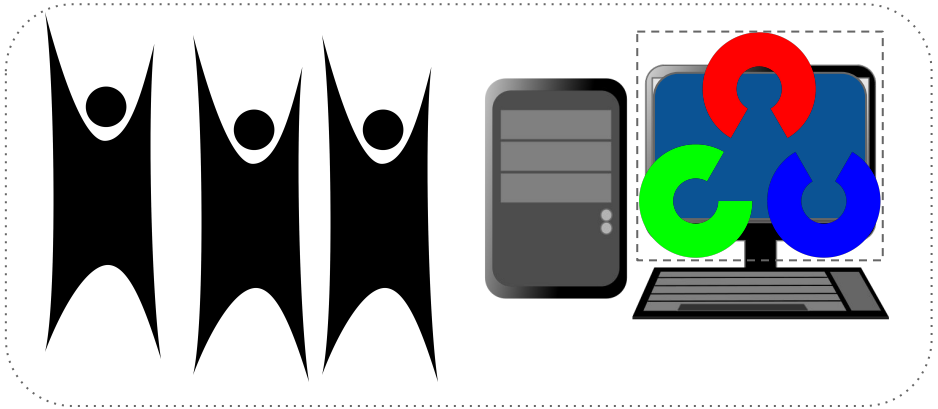


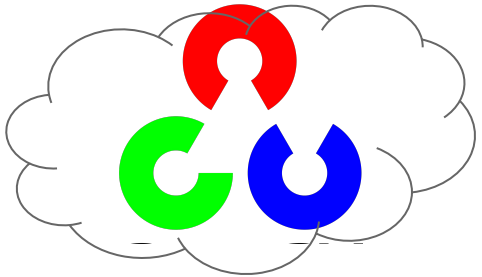
Towards Distributed Build System



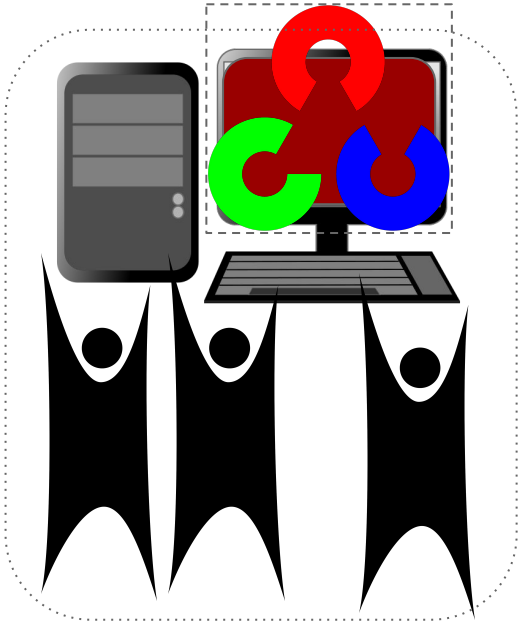
Towards Distributed Build System



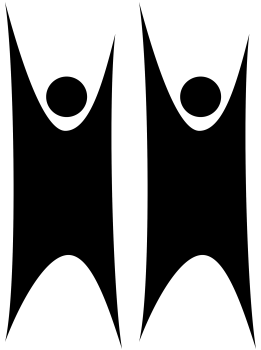


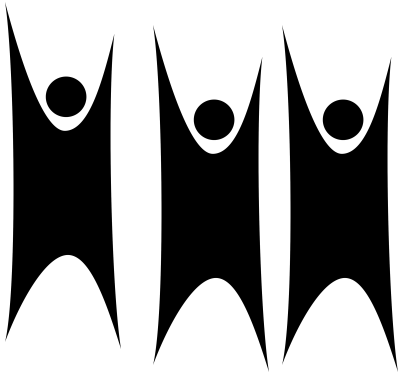


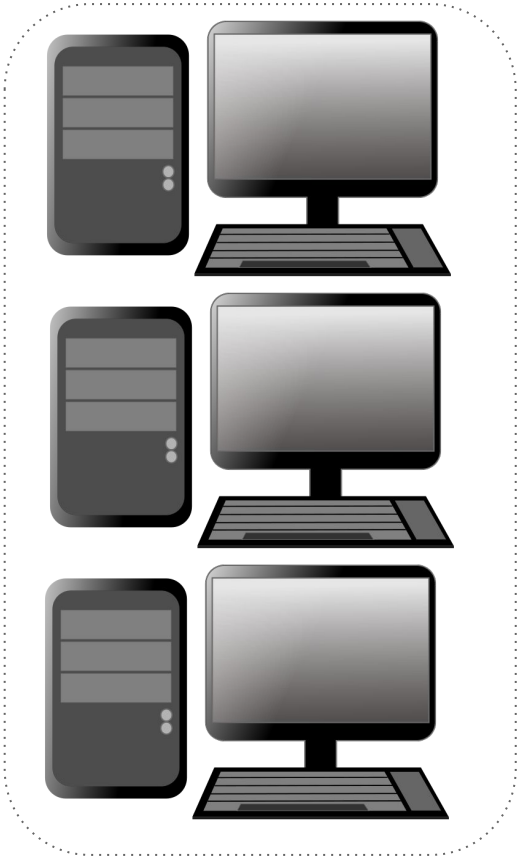
http://opencv.org/

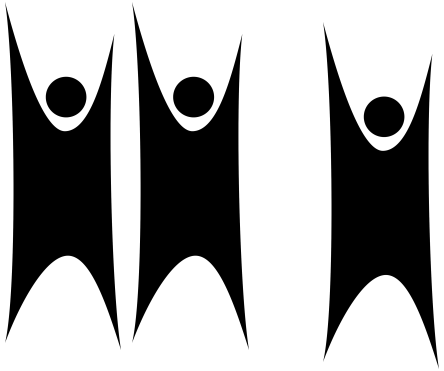


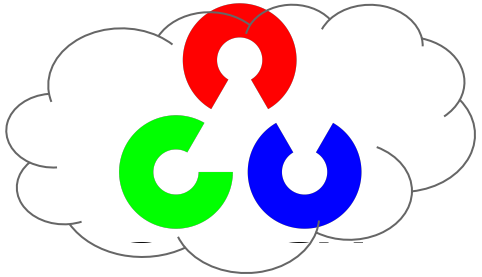
Towards Distributed Build System





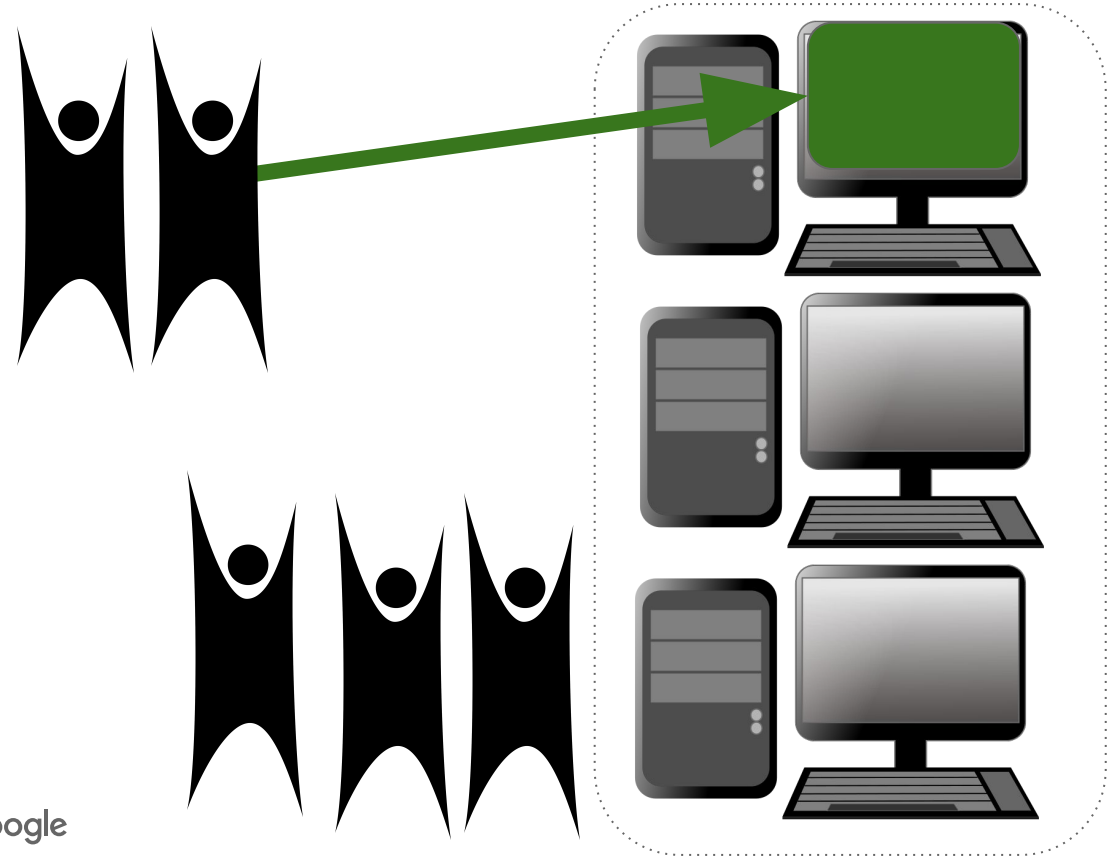


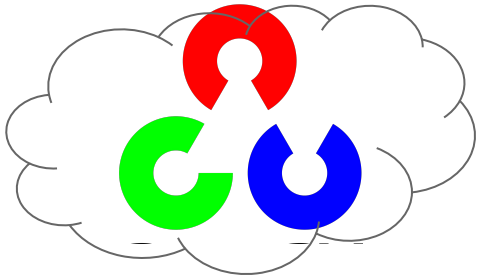




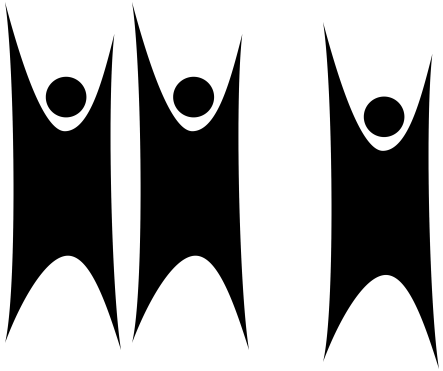
http://opencv.org/

Towards Distributed Build System

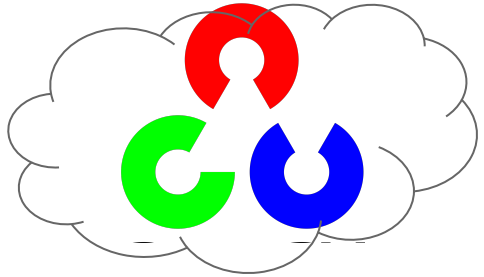




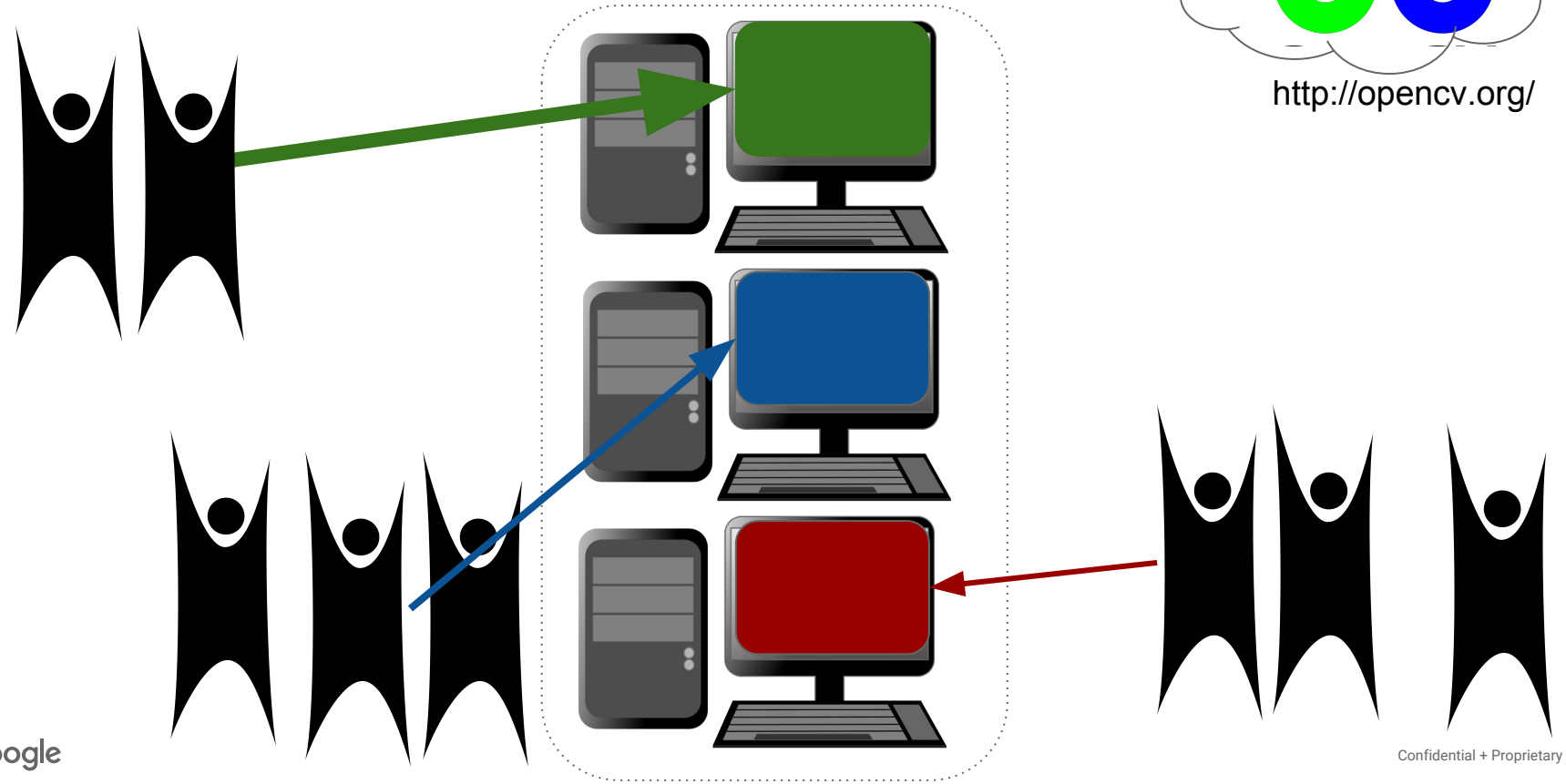
http://opencv.org/



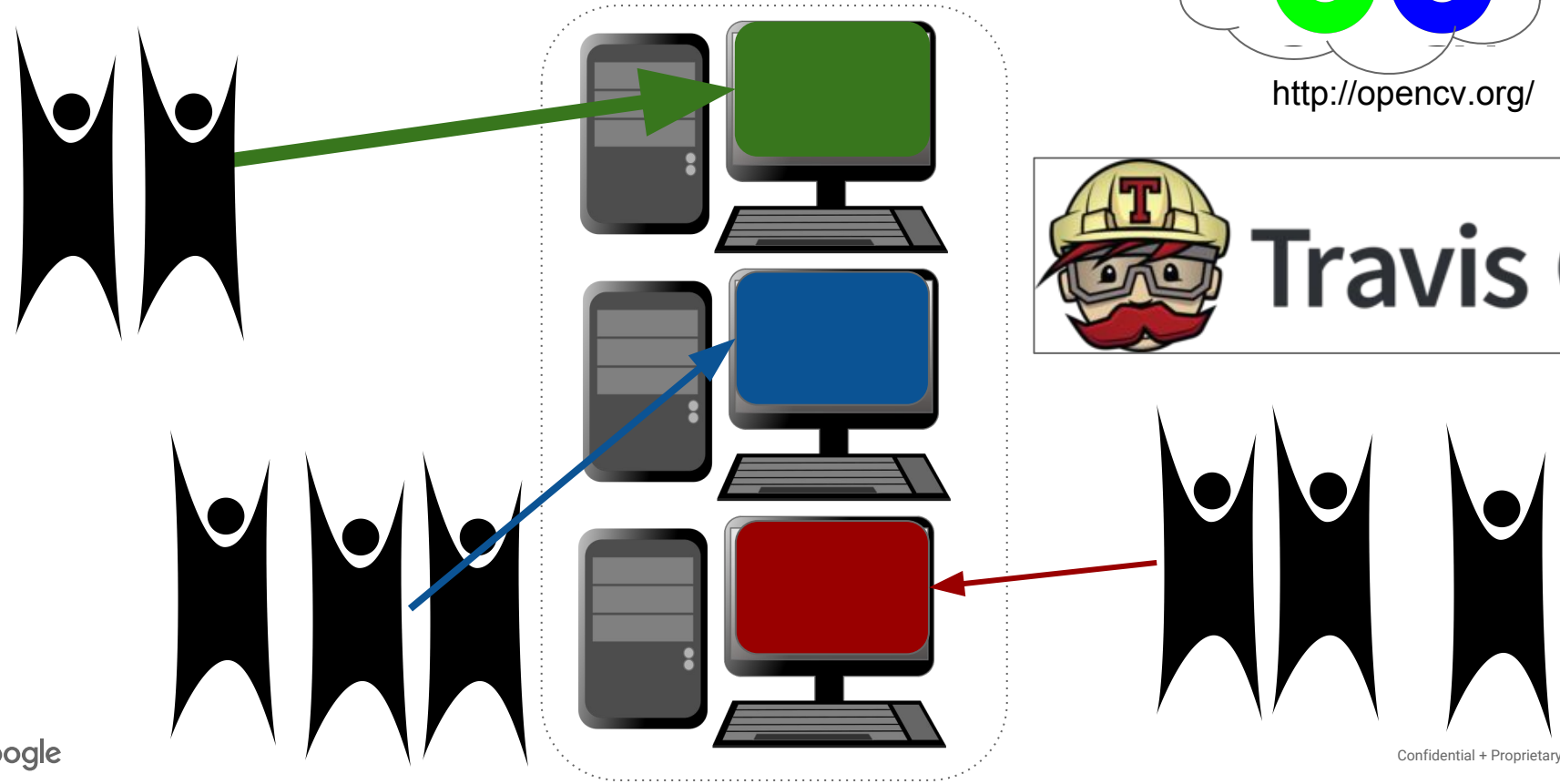
Towards Distributed Build System



http://opencv.org/



Towards Distributed Build System



http://opencv.org/

Travis CI

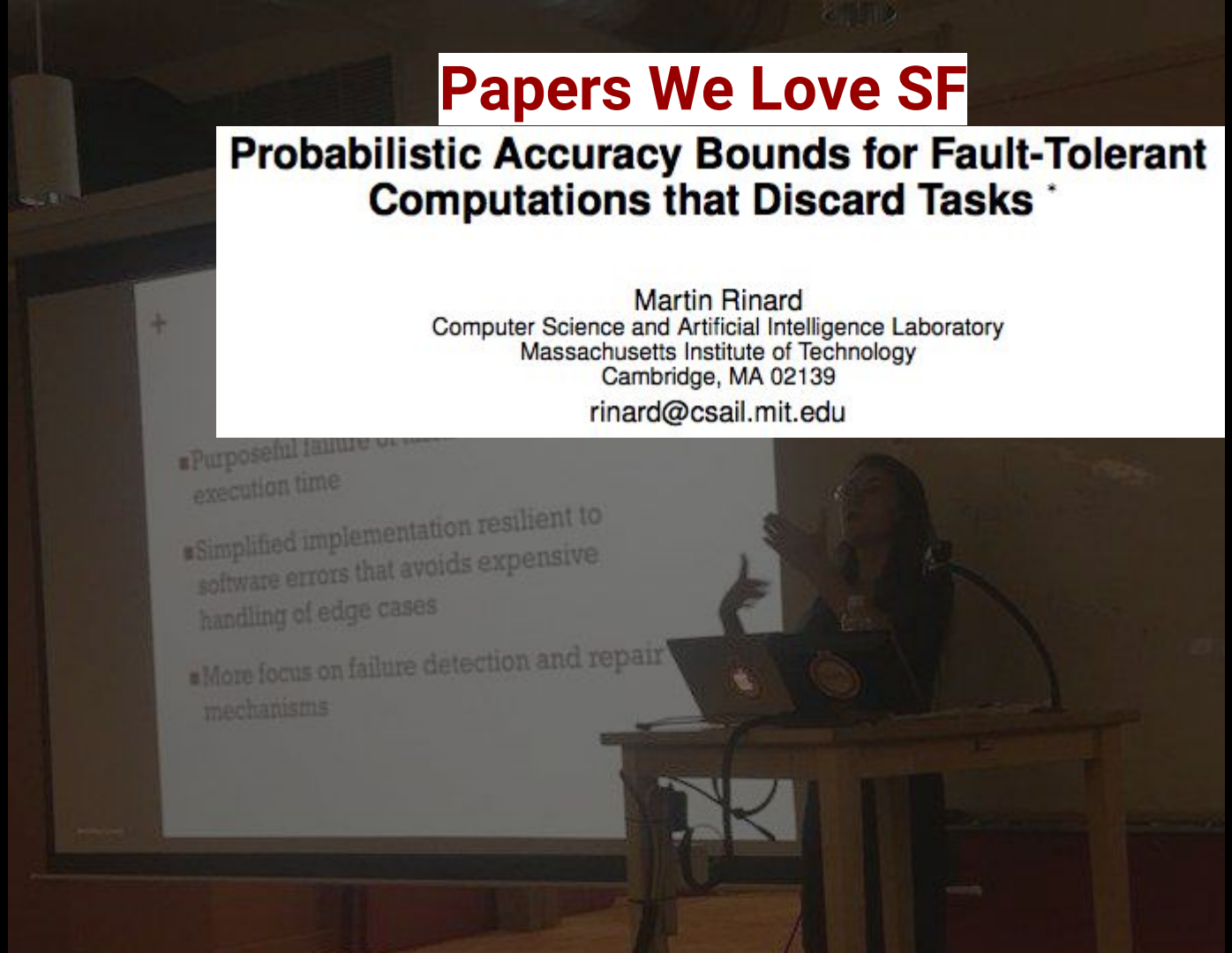
Papers We Love NYC

One VM to Rule Them All

Thomas Würthinger* Christian Wimmer* Andreas Wöß† Lukas Stadler†
Gilles Duboscq† Christian Humer† Gregor Richards§ Doug Simon* Mario Wolczko*

*Oracle Labs †Institute for System Software, Johannes Kepler University Linz, Austria §S³ Lab, Purdue University
{thomas.wuerthinger, christian.wimmer, doug.simon, mario.wolczko}@oracle.com
{woess, stadler, duboscq, christian.humer}@ssw.jku.at gr@purdue.edu

Papers We Love SF



Aysylu Greenberg

Google

$f(x) = x$

@aysylu22

Papers We Love - London

Home | Members | Photos | Discussions | More

$f(x) = x$

Reading and Talking about Papers.

+ SUGGEST A NEW MEETUP

Upcoming 1 | Past | Calendar

Fouad Mardini on Christofides on The Travelling Salesman

Skills Matter at CodeNode
10 South Place, London (map)

Wed Mar 16
6:30 PM

London, United

meetup.com/Papers-We-Love-London

Today

- Staged Event-Driven Architecture

Today

- Staged Event-Driven Architecture
- Leases

Today

- Staged Event-Driven Architecture
- Leases
- Inaccurate Computations

Computer Science Research In Distributed Systems Industry

Operating systems research

AN EXPERIMENTAL TIME-SHARING SYSTEM

Fernando J. Corbató, Marjorie Merwin Daggett, Robert C. Daley

Computation Center, Massachusetts Institute of Technology

Operating systems research

AN EXPERIMENTAL TIME-SHARING SYSTEM

Fernando J. Corbató, Marjorie Merwin Daggett, Robert C. Daley

Computation Center, Massachusetts Institute of Technology

Operating systems research

Concurrency

COOPERATING
SEQUENTIAL PROCESSES

EDSGER W. DIJKSTRA

(1965)

Operating systems research

Concurrency

COOPERATING
SEQUENTIAL PROCESSES

EDSGER W. DIJKSTRA

(1965)

Concurrency primitives: mutex & semaphore

Operating systems research

Concurrency

COOPERATING
SEQUENTIAL PROCESSES

Processes execute at different speeds

Concurrency primitives: mutex & semaphore

Time in distributed systems

Time, Clocks, and the Ordering of Events in a Distributed System

Leslie Lamport
Massachusetts Computer Associates, Inc.

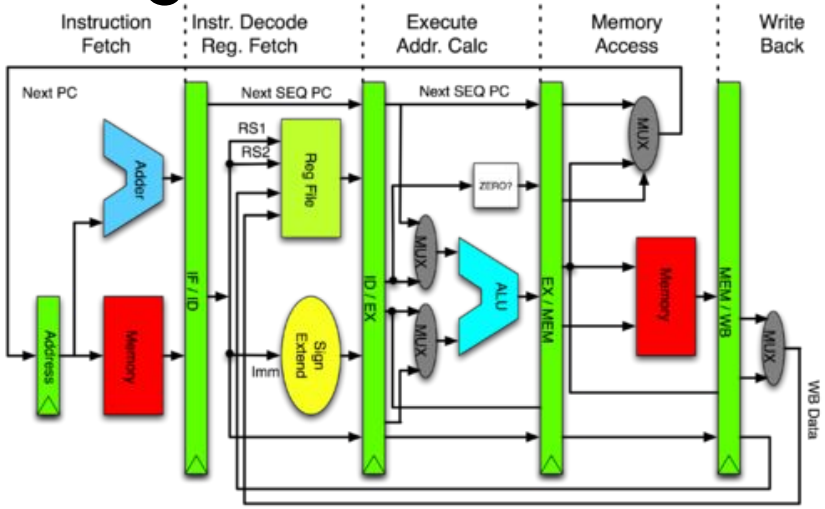
Time in distributed systems

Time, Clocks, and the Ordering of Events in a Distributed System

Leslie Lamport
Massachusetts Computer Associates, Inc.

Time in distributed systems

Pipelining



Time in distributed systems

Pipelining

Internet

Internet

Distributed consensus

Internet

Distributed consensus

Viewstamped Replication: A New Primary Copy Method to Support Highly-Available Distributed Systems

Brian M. Oki
Barbara H. Liskov

Massachusetts Institute of Technology

Internet

Distributed consensus

Viewstamped Replication: A New Primary Copy Method to Support Highly-Available Distributed Systems

Brian M. Oki
Barbara H. Liskov

Massachusetts Institute of Technology

The Part-Time Parliament

Leslie Lamport

Internet

Distributed consensus

Viewstamped Replication: A New Primary Copy Method to Support Highly-Available Distributed Systems

Brian M. Oki
Barbara H. Liskov

Massachusetts Institute of Technology

The Part-Time Parliament

Leslie Lamport

Paxos

Reconsider large systems

Reconsider large systems

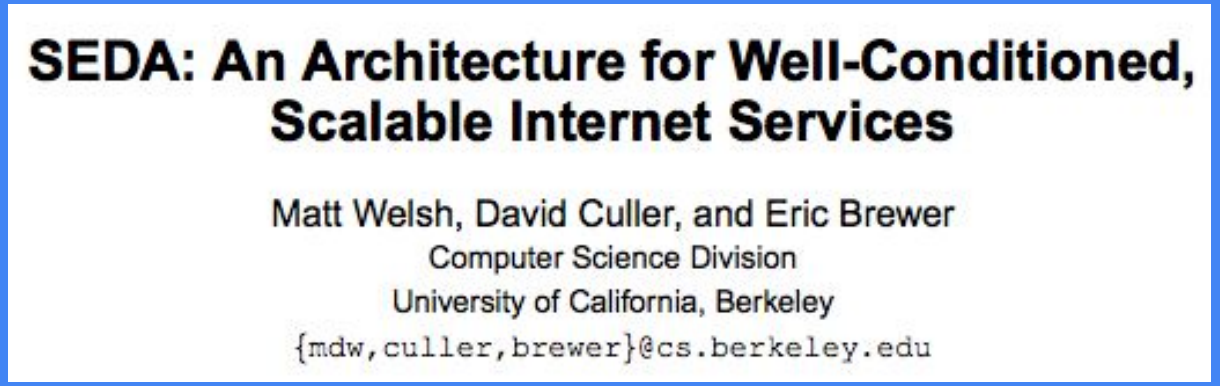
Platform as a service

CS Research is Timeless

Inform decisions

Mitigate technical risk

Staged Event Driven Architecture & *Deep Pipelines*

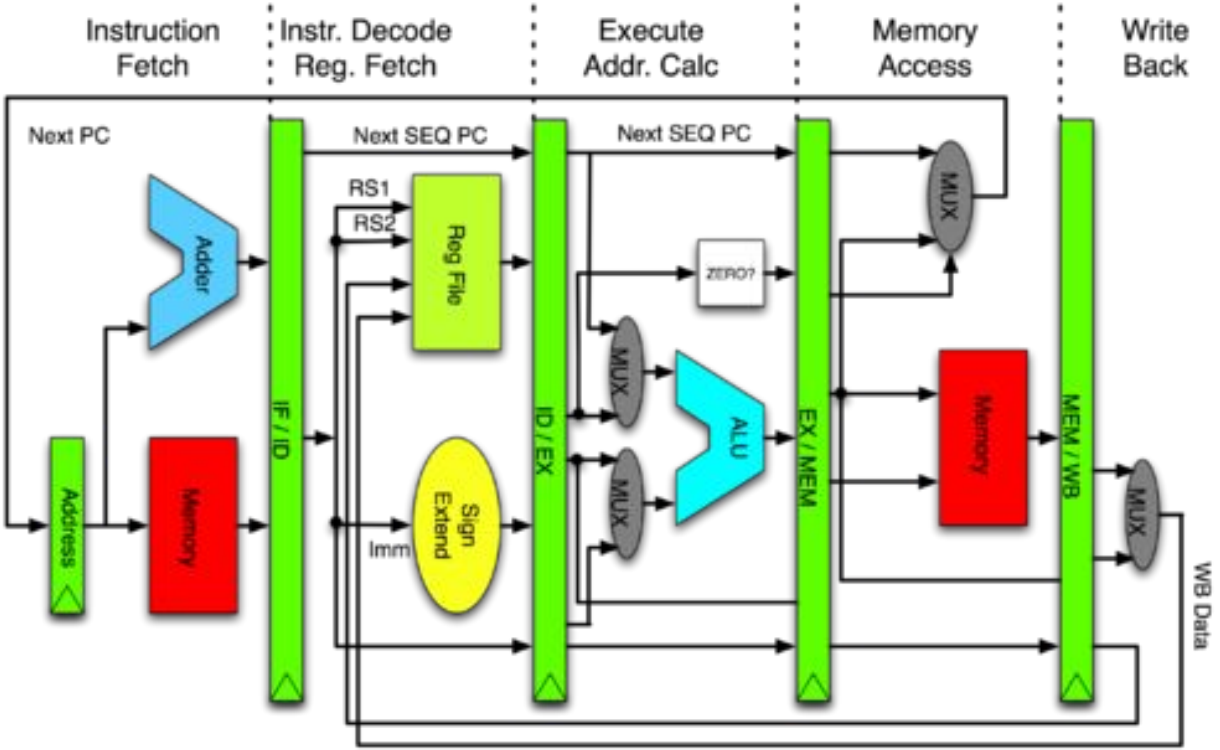


SEDA: An Architecture for Well-Conditioned, Scalable Internet Services

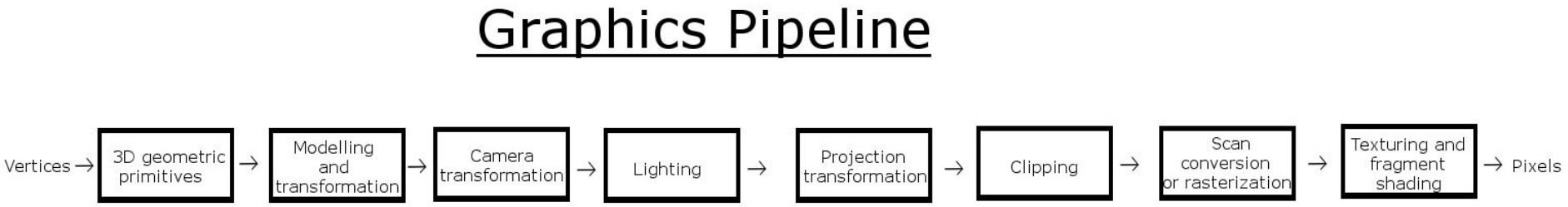
Matt Welsh, David Culler, and Eric Brewer
Computer Science Division
University of California, Berkeley
{mdw,culler,brewer}@cs.berkeley.edu

2001

Hardware Pipelines

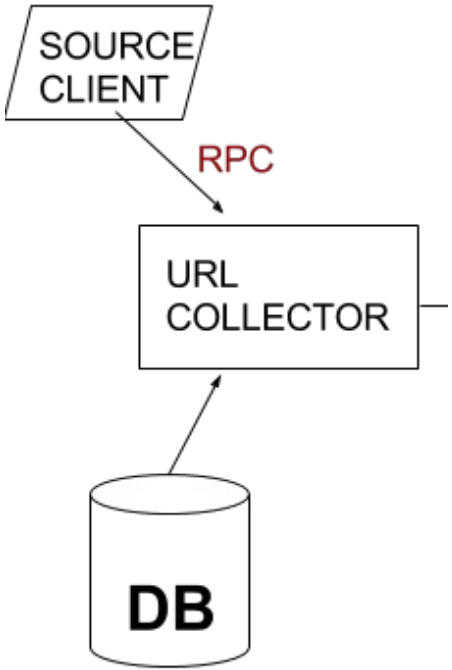


Data Pipelines

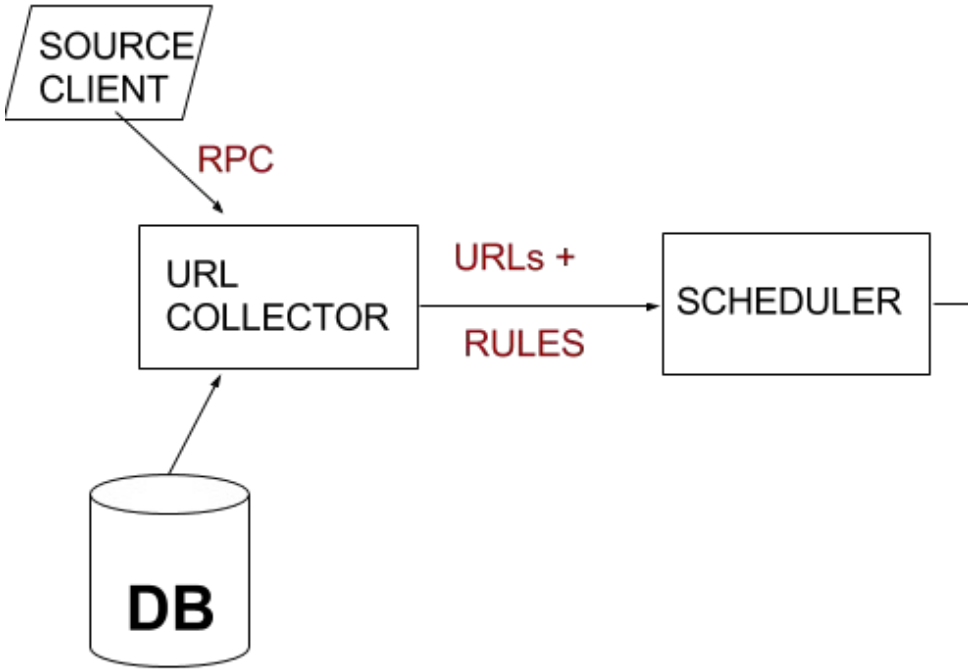


https://en.wikipedia.org/wiki/Graphics_pipeline

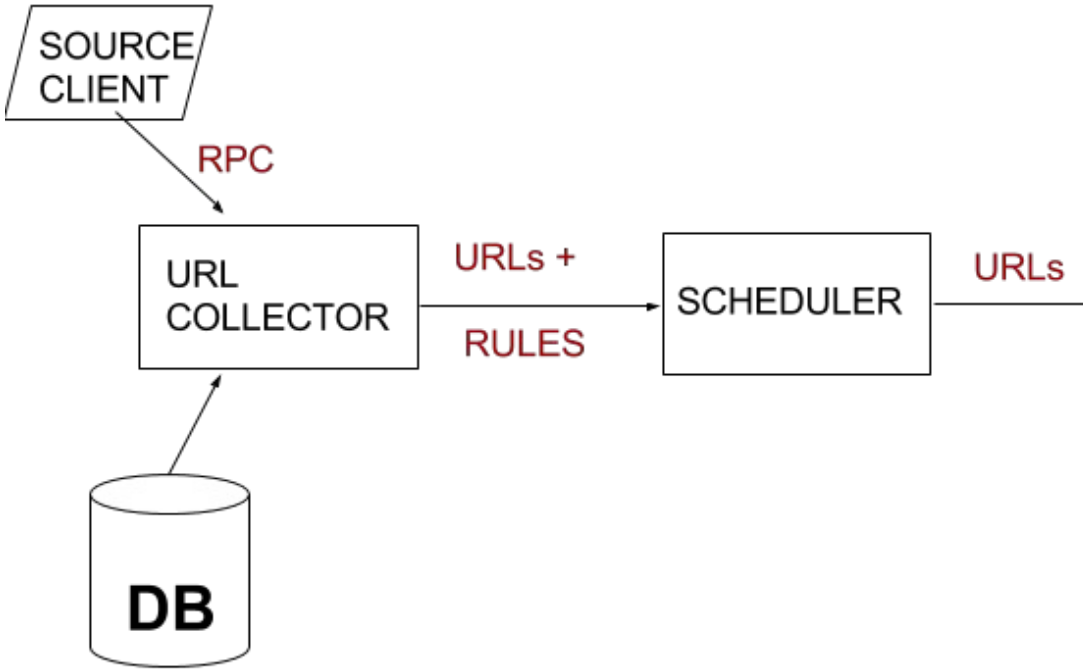
Search Indexing Pipelines



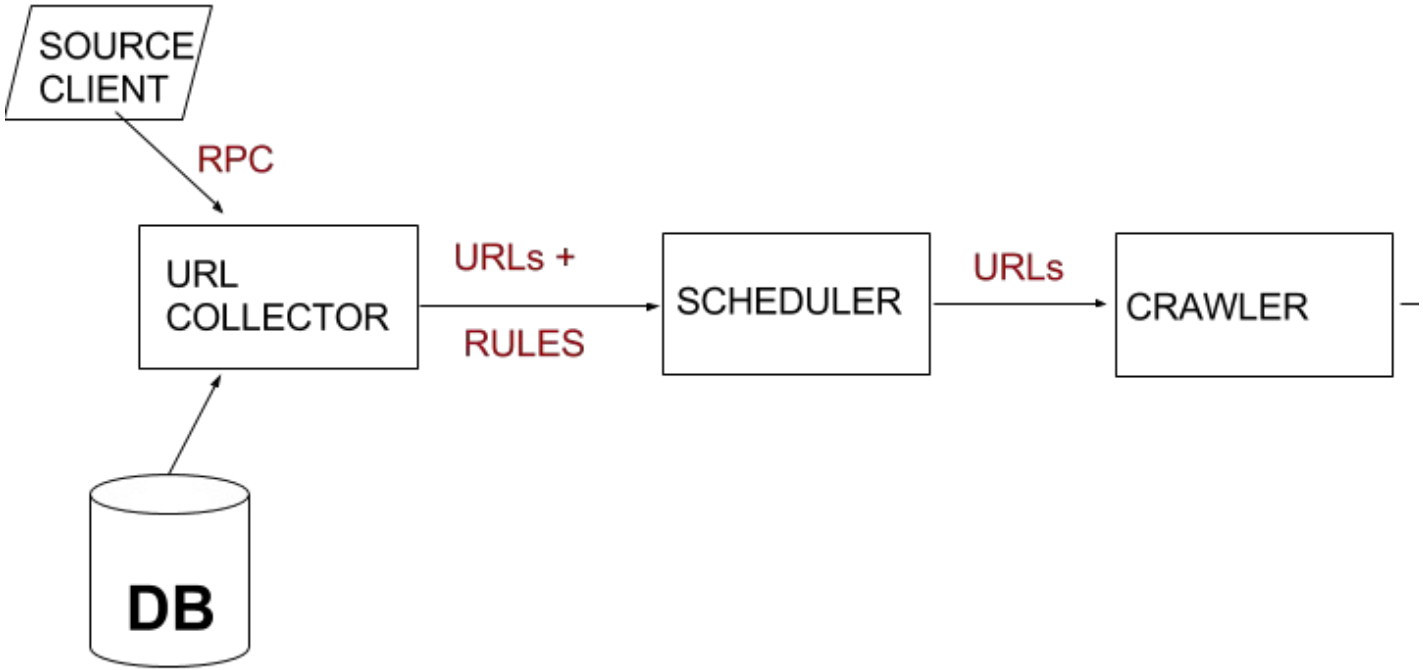
Search Indexing Pipelines



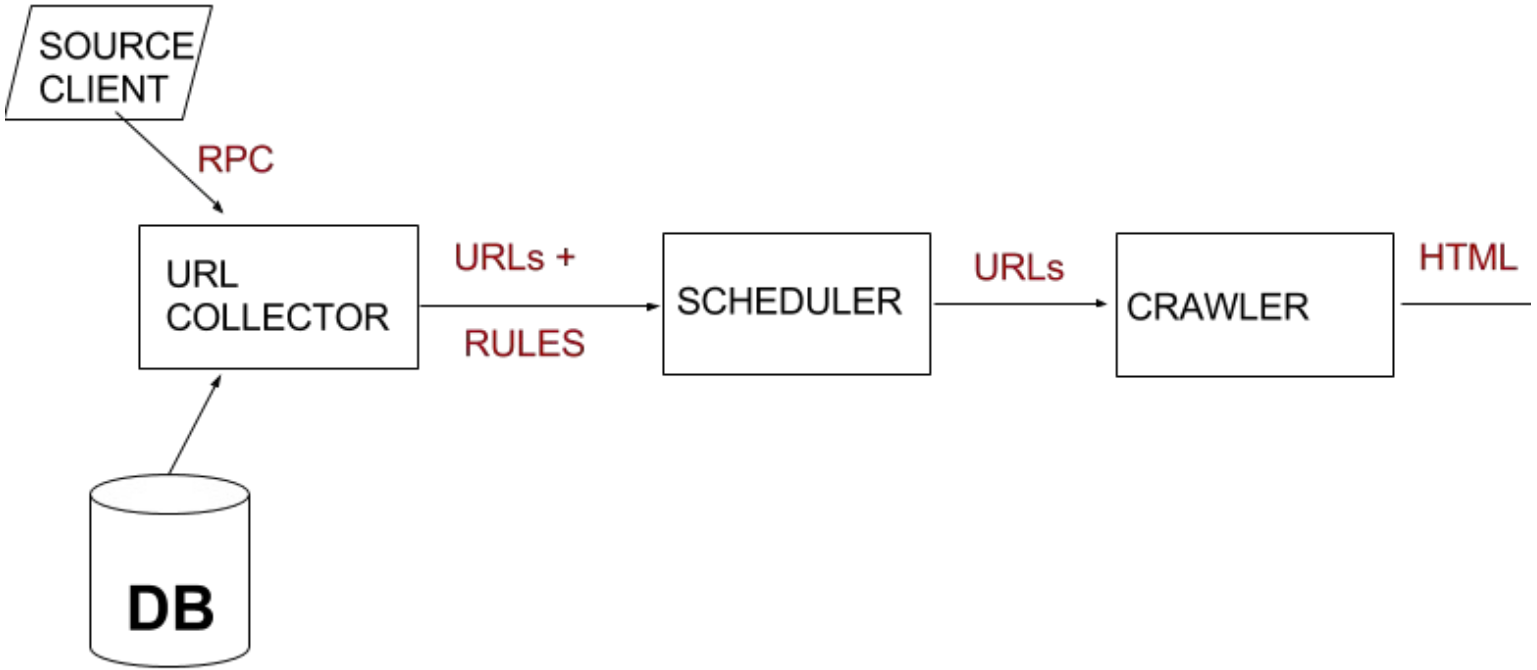
Search Indexing Pipelines



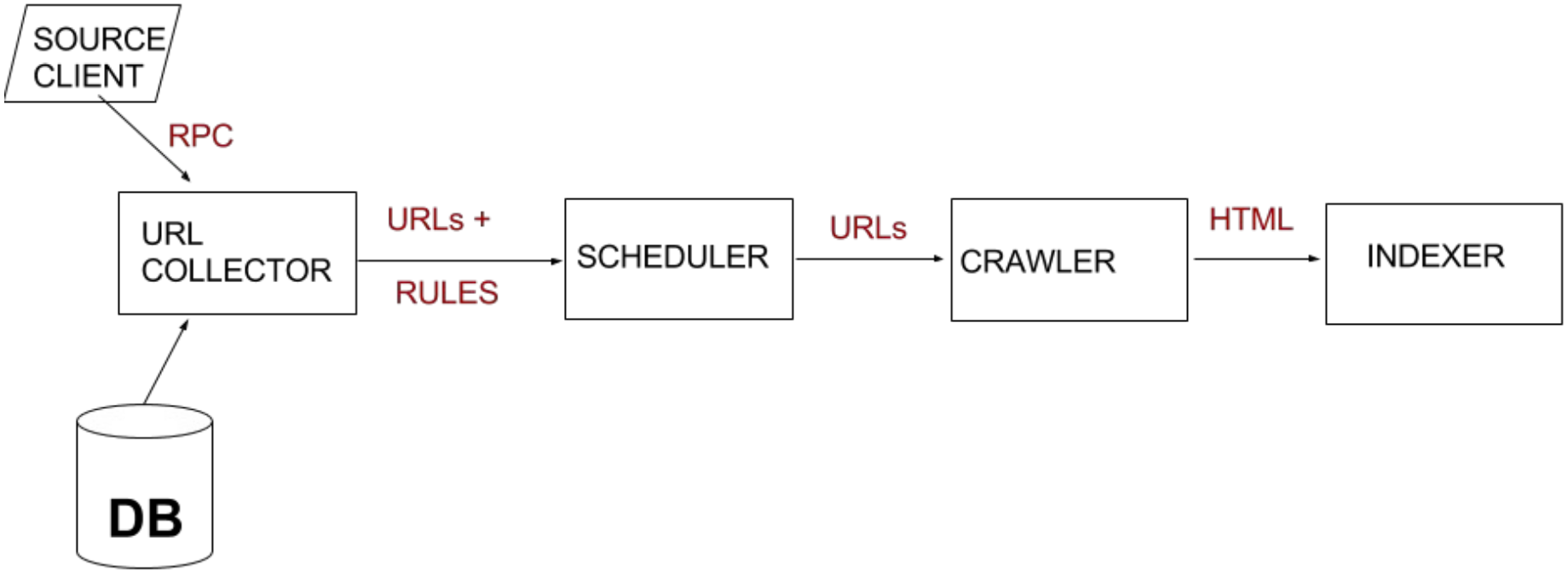
Search Indexing Pipelines



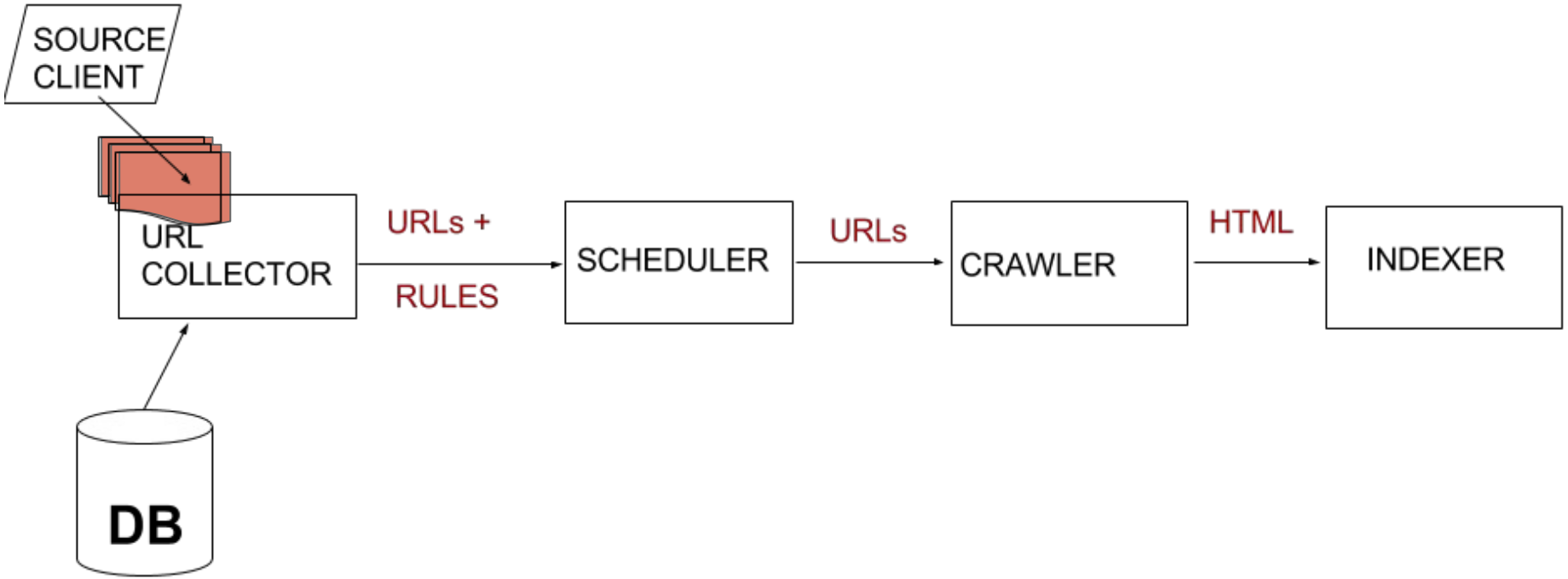
Search Indexing Pipelines



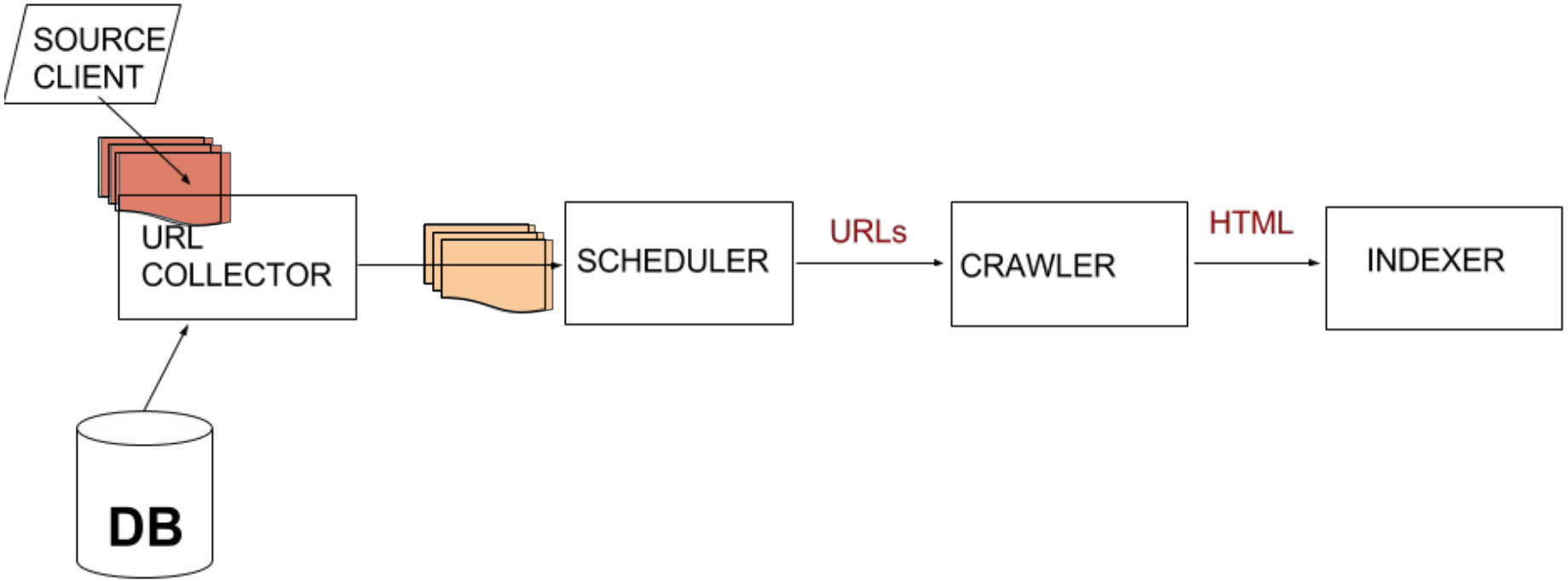
Search Indexing Pipelines



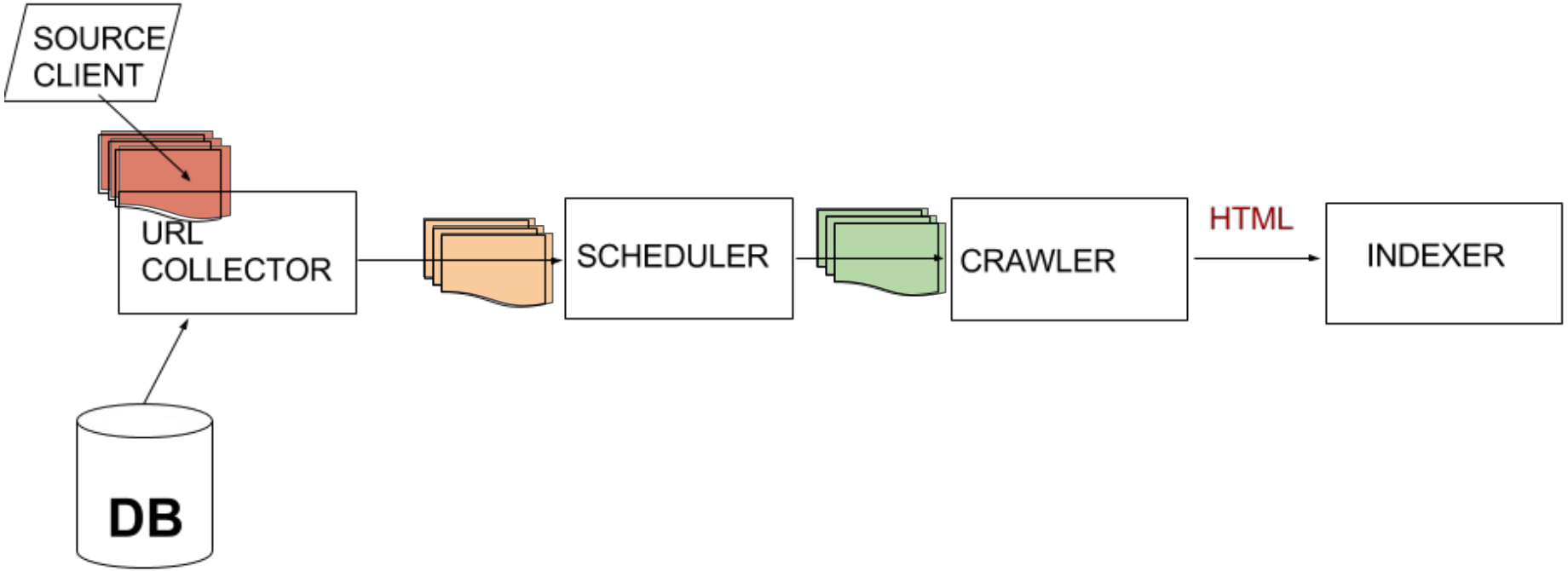
Search Indexing Pipelines



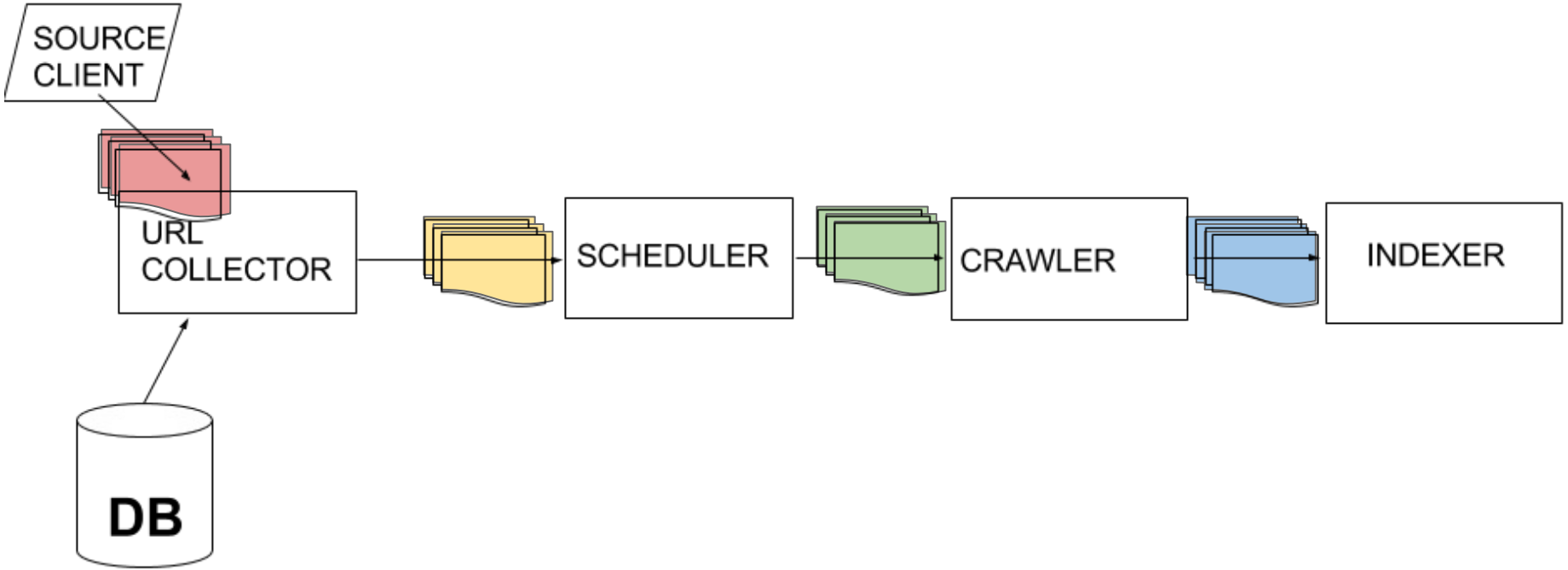
Search Indexing Pipelines



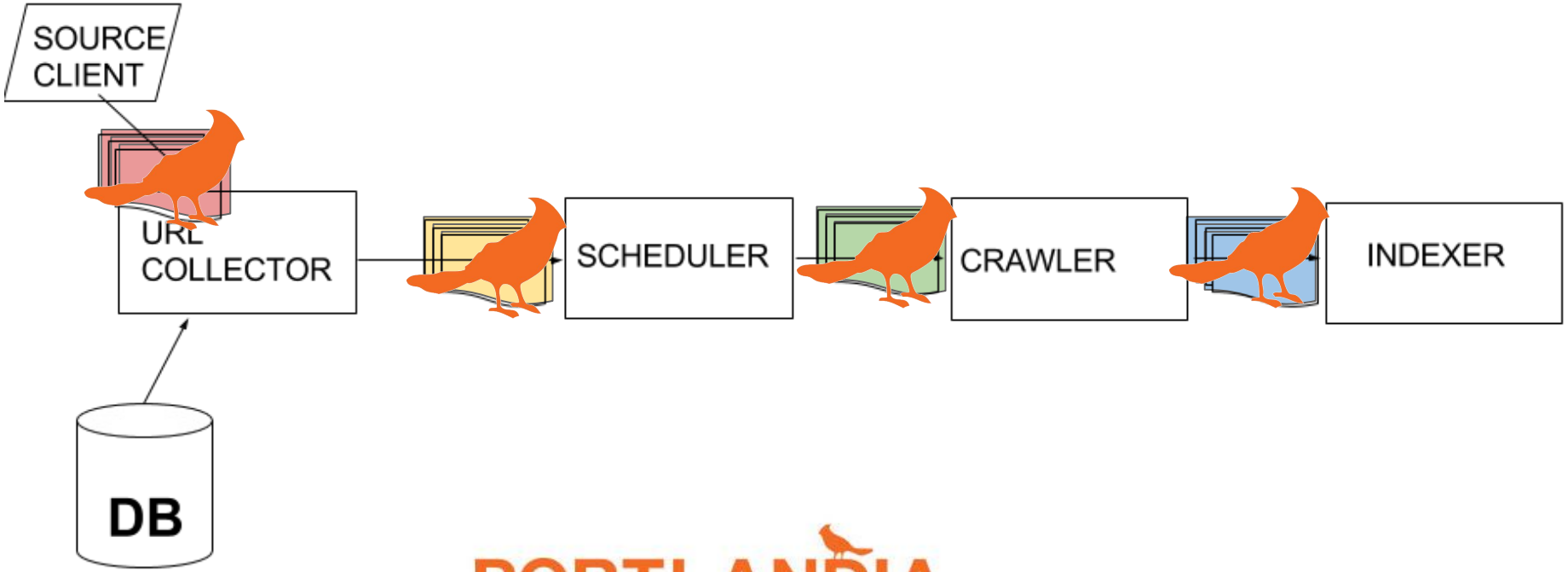
Search Indexing Pipelines



Search Indexing Pipelines



Search Indexing Pipelines

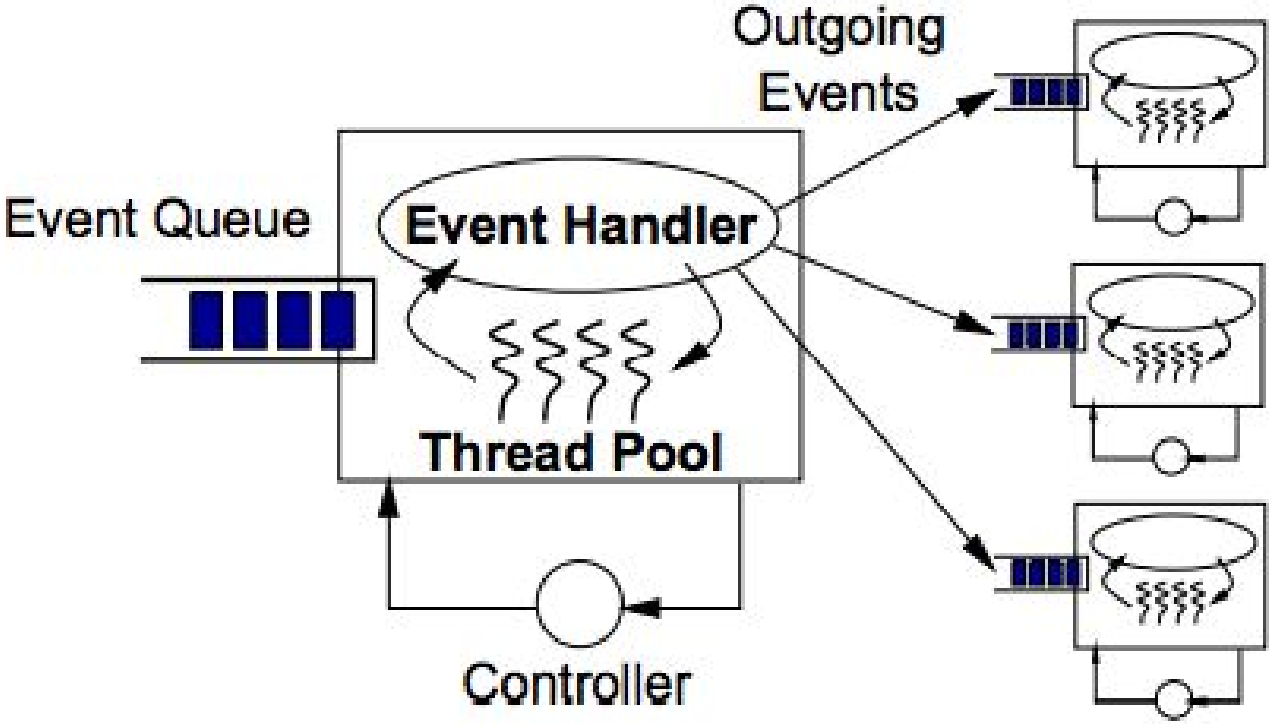


PORTLANDIA

SEDA: An Architecture for Well-Conditioned, Scalable Internet Services

Matt Welsh, David Culler, and Eric Brewer
Computer Science Division
University of California, Berkeley
{mdw,culler,brewer}@cs.berkeley.edu

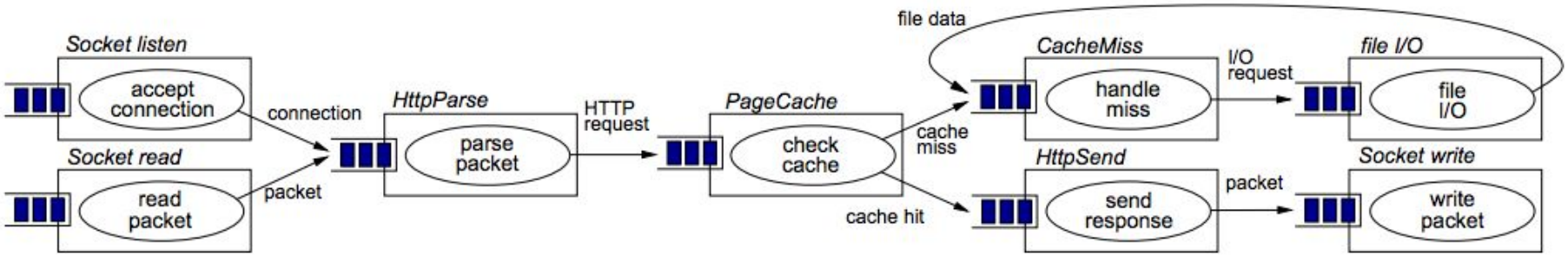
Staged Event Driven Architecture

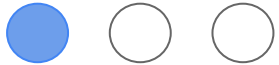


Staged Event Driven Architecture

Single-machine pipeline

generalizes to distributed pipelines



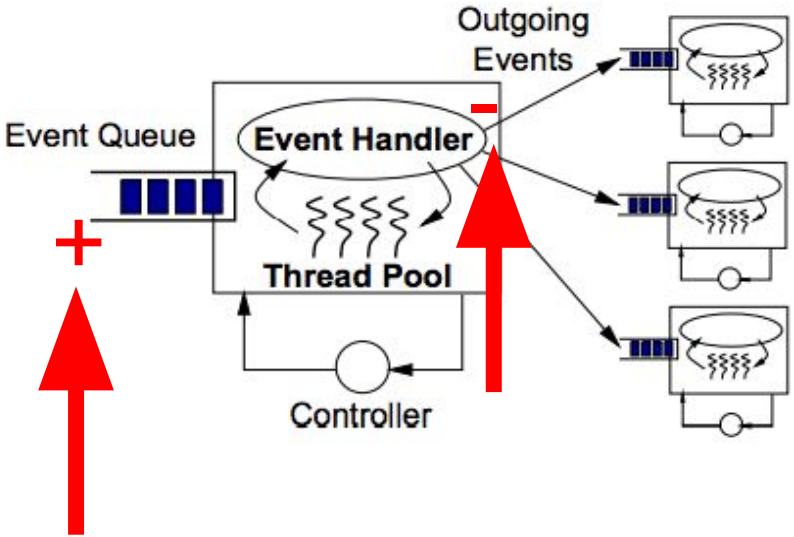


Staged Event Driven Architecture

- Dynamic resource controllers

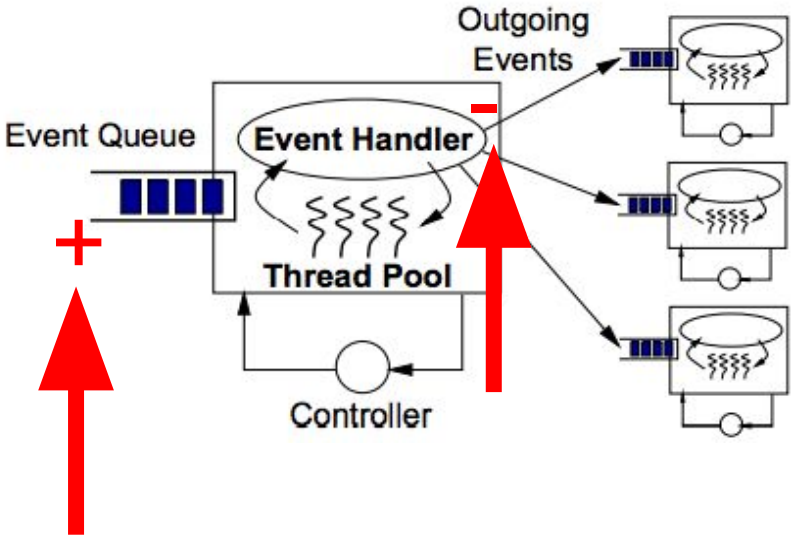
Staged Event Driven Architecture

- Dynamic resource controllers



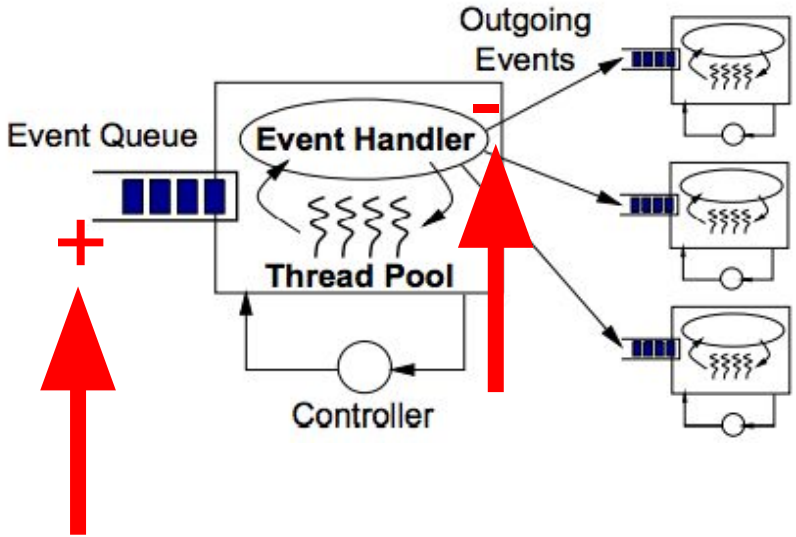
Staged Event Driven Architecture

- Dynamic resource controllers
  - automatic tuning



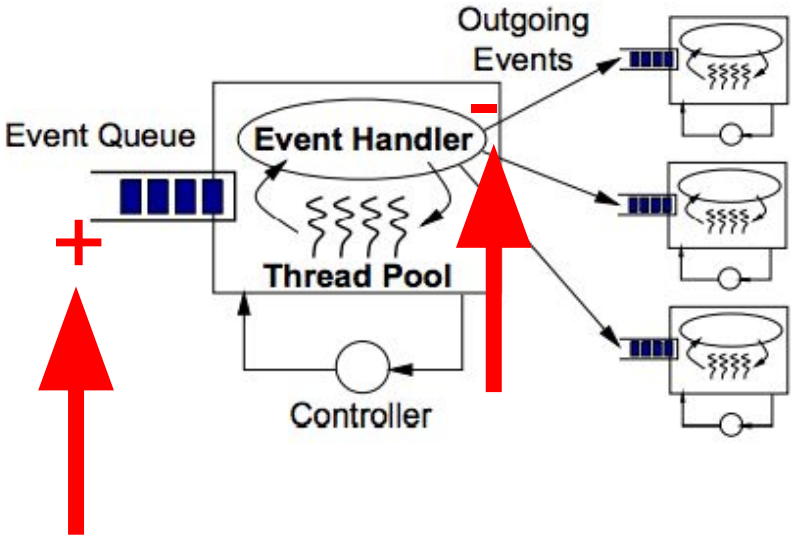
Staged Event Driven Architecture

- Dynamic resource controllers
  - automatic tuning
    - thread pool sizing



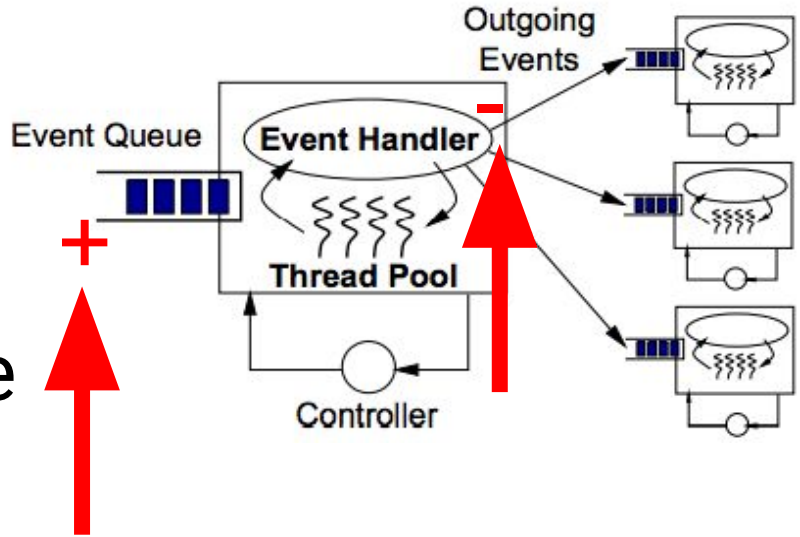
Staged Event Driven Architecture

- Dynamic resource controllers
  - automatic tuning
    - thread pool sizing
    - event batching



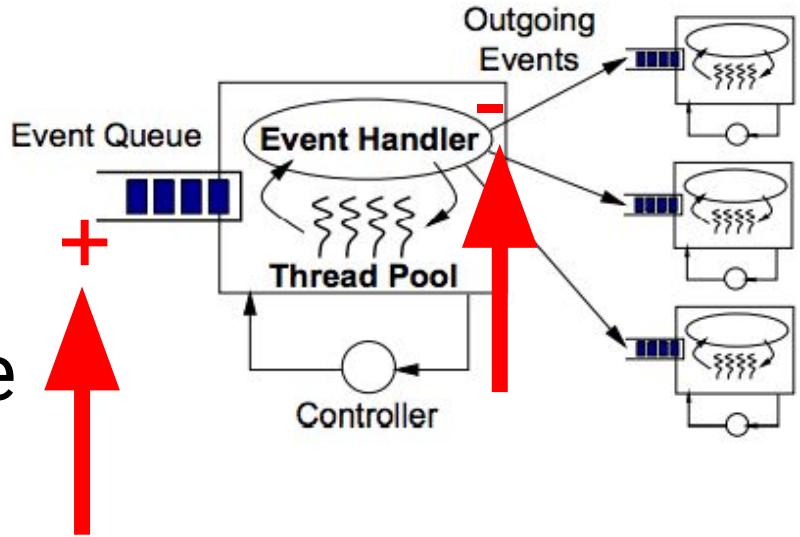
Staged Event Driven Architecture

- Dynamic resource controllers
  - automatic tuning
    - thread pool sizing
    - event batching
  - load shedding via queue



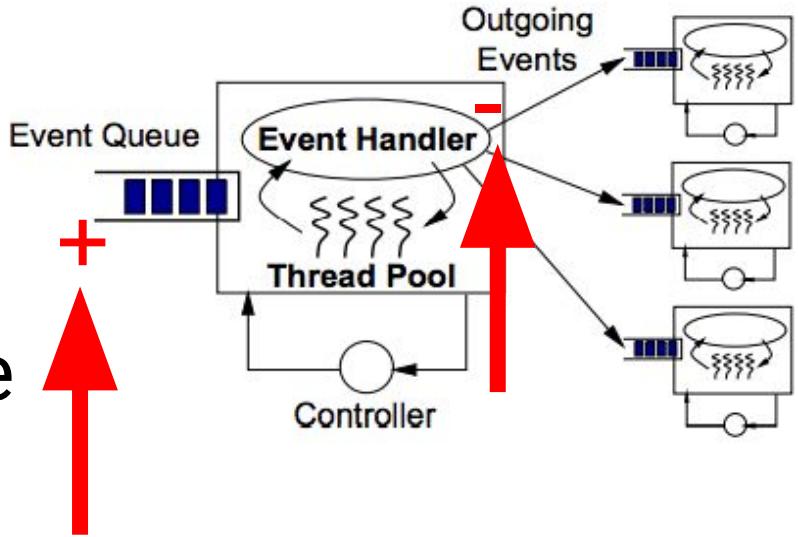
Staged Event Driven Architecture

- Dynamic resource controllers
  - automatic tuning
    - thread pool sizing
    - event batching
  - load shedding via queue
    - backpressure

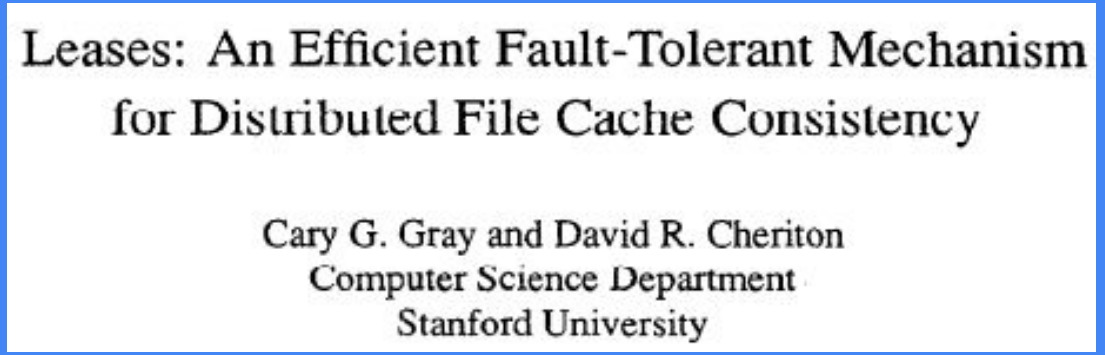


Staged Event Driven Architecture

- Dynamic resource controllers
  - automatic tuning
    - thread pool sizing
    - event batching
  - load shedding via queue
    - backpressure
    - threshold, filter, reorder, aggregate



Leases as Heart Beat in Distributed Systems


Leases: An Efficient Fault-Tolerant Mechanism for Distributed File Cache Consistency

Cary G. Gray and David R. Cheriton
Computer Science Department
Stanford University

1989

Leases: An Efficient Fault-Tolerant Mechanism for Distributed File Cache Consistency

Cary G. Gray and David R. Cheriton
Computer Science Department
Stanford University

Leases

- Distributed locking

Leases

- Distributed locking
- Lease term tradeoffs
  - short

Leases

- Distributed locking
- Lease term tradeoffs
  - short
    - delay from client and server failures minimized
    - reduced false write-sharing

Leases

- Distributed locking
- Lease term tradeoffs
  - short vs long

Leases

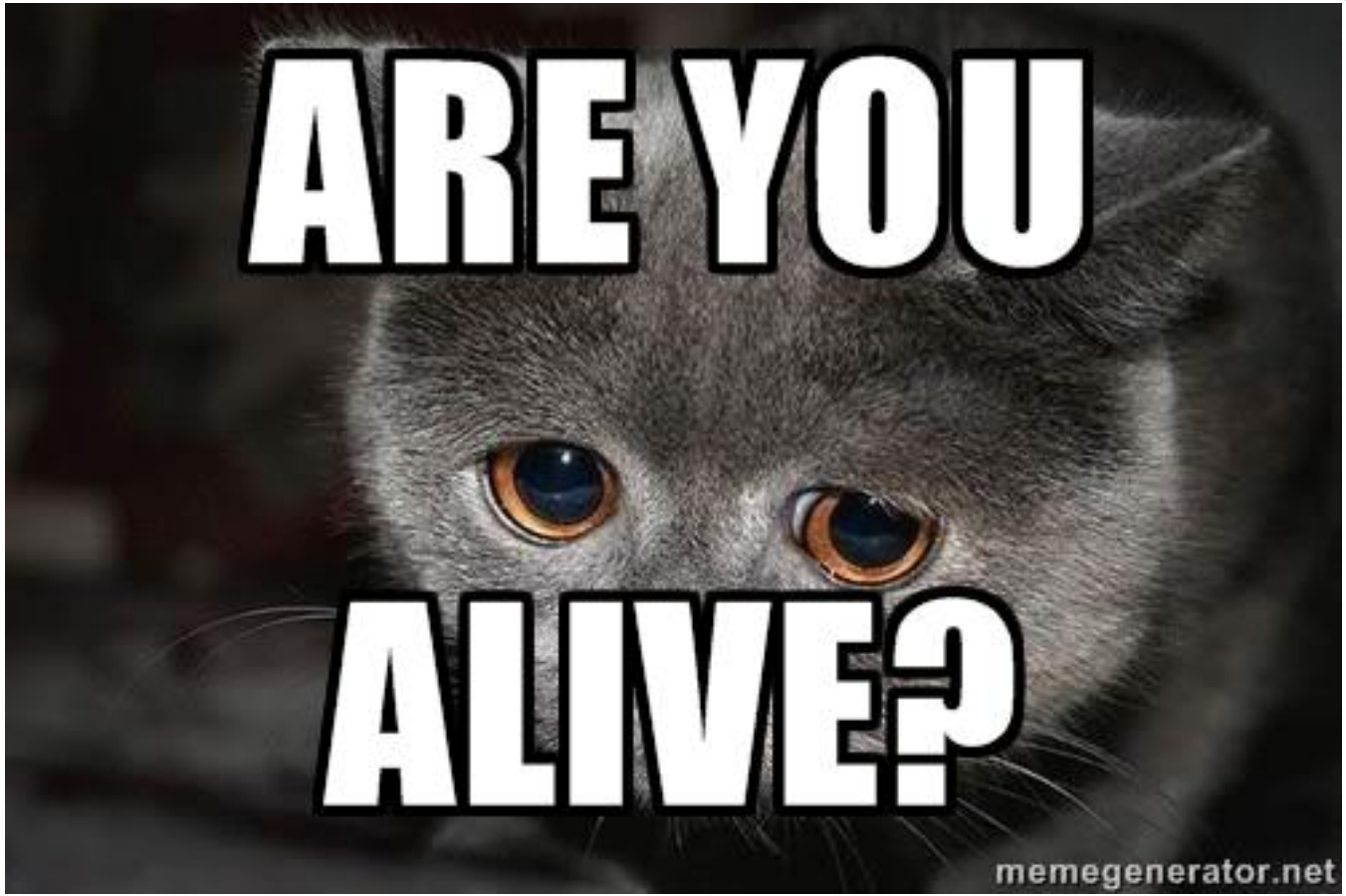
- Distributed locking
- Lease term tradeoffs
  - short vs long
    - more efficient for frequently accessed data
    - minimized lease extension overhead on server and client

Leases

- Distributed locking
- Lease term tradeoffs
  - short vs long
- Use of leases in modern applications
  - Leader election TTL (in etcd)

Leases

- Distributed locking
- Lease term tradeoffs
  - short vs long
- Use of leases in modern applications
  - Leader election TTL (in etcd)
  - Liveness detection



Do the build, s'il vous plaît!

Build System

Dita Roque-Gourary

Google

Do the build, s'il vous plaît!

OK/D'Accord!

Build System

Dita Roque-Gourary

Google

Do the build, s'il vous plaît!

OK/D'Accord!

Waiting for the results, merci.

Build System

Dita Roque-Gourary

Google

Do the build, s'il vous plaît!

OK/D'Accord!

Waiting for the results, merci.

Build is in progress

Build System

Dita Roque-Gourary

Google

Dita Roque-Gourary

Google

Do the build, s'il vous plaît!

OK/D'Accord!

Waiting for the results, merci.

Build is in progress

Waiting for the results, merci.

Build System

Do the build, s'il vous plaît!

OK/D'Accord!

Waiting for the results, merci.

Build is in progress

Waiting for the results, merci.

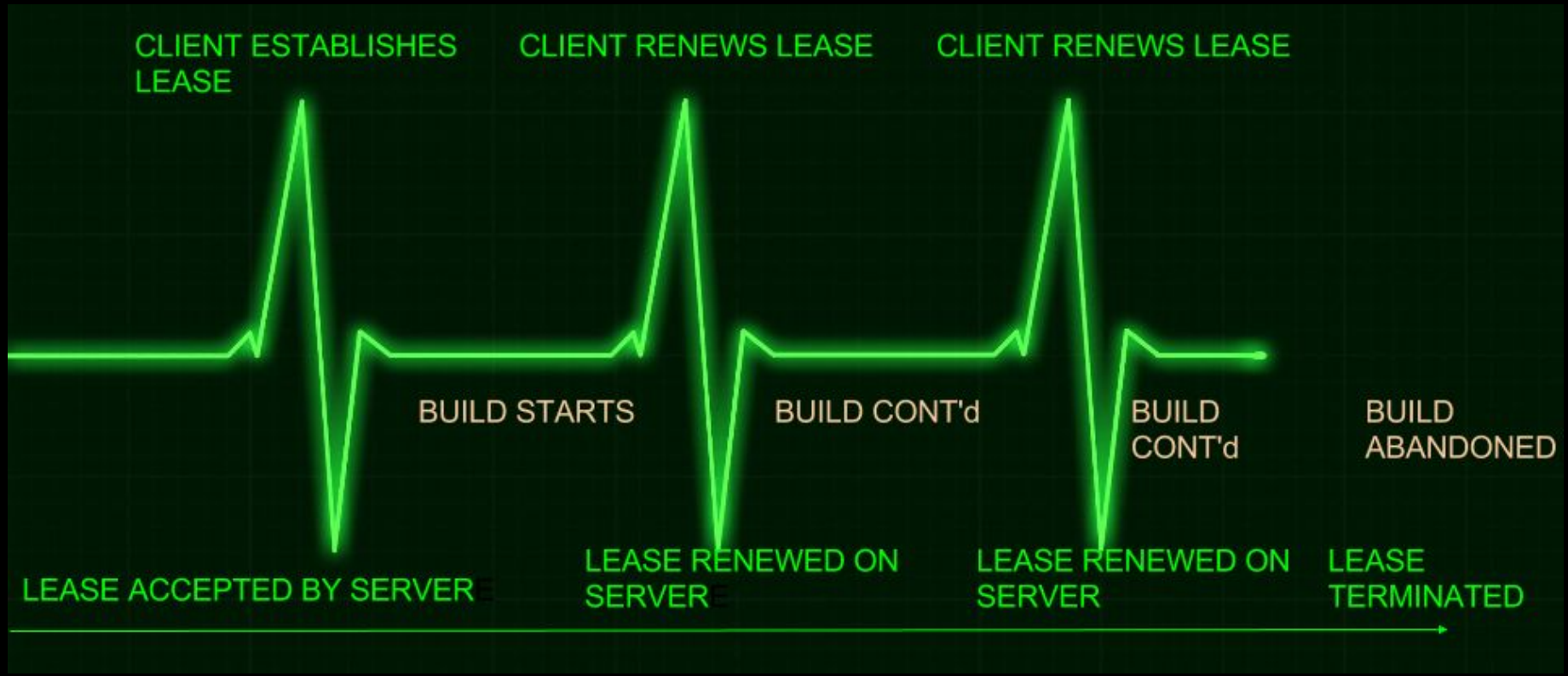
Build is finished/fini.

Build System

Dita Roque-Gourary

Google

Leases in Build System



Inaccurate Computations & Serving Search Results

From Accurate to "Good Enough"

```
SELECT COUNT(*)
FROM Sessions
WHERE Genre = 'western'
GROUP BY OS
ERROR WITHIN 10% AT CONFIDENCE 95%
```

```
SELECT COUNT(*)
FROM Sessions
WHERE Genre = 'western'
GROUP BY OS
ERROR WITHIN 10% AT CONFIDENCE 95%
```

```
SELECT COUNT(*)
FROM Sessions
WHERE Genre = 'western'
GROUP BY OS
WITHIN 5 SECONDS
```

BlinkDB

BlinkDB: Queries with Bounded Errors and Bounded Response Times on Very Large Data

Sameer Agarwal[†], Barzan Mozafari[°], Aurojit Panda[†], Henry Milner[†], Samuel Madden[°], Ion Stoica[*†]

Probabilistic Accuracy Bounds for Fault-Tolerant Computations that Discard Tasks *

Martin Rinard
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139
rinard@csail.mit.edu

Inaccuracy for Resilience

1. Task decomposition

Inaccuracy for Resilience

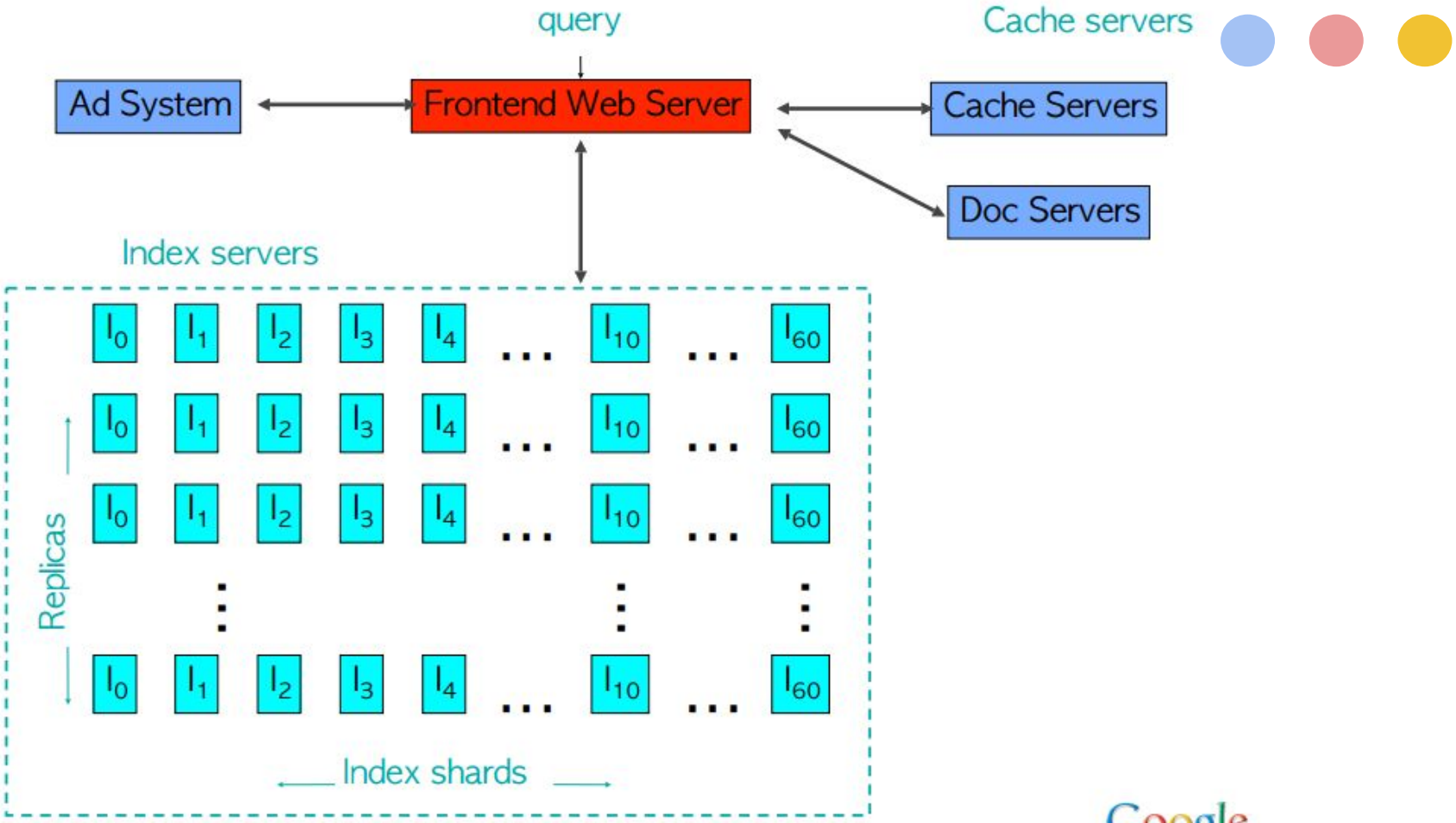
1. Task decomposition
2. Baseline for correctness

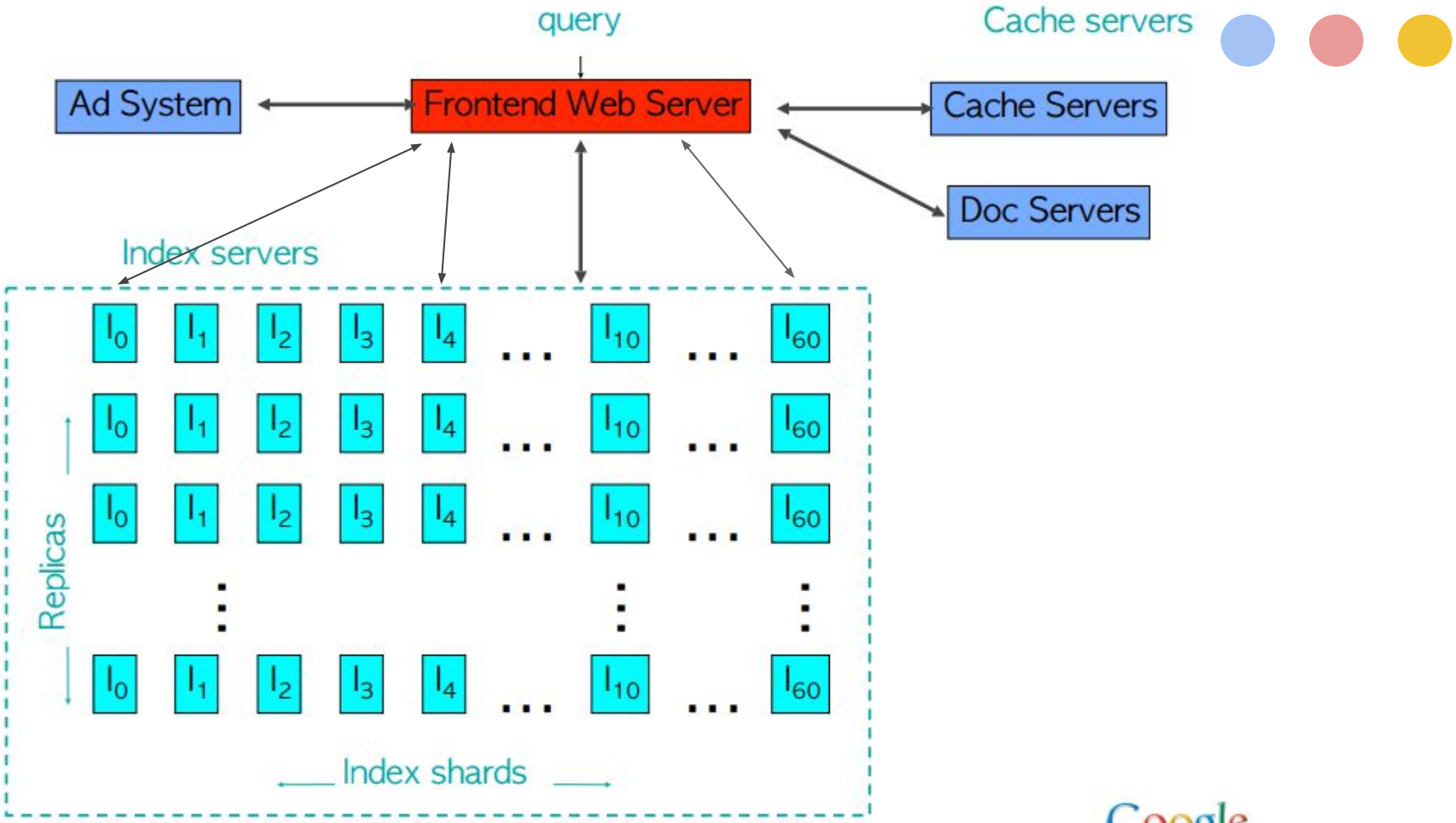
Inaccuracy for Resilience

1. Task decomposition
2. Baseline for correctness
3. Criticality Testing

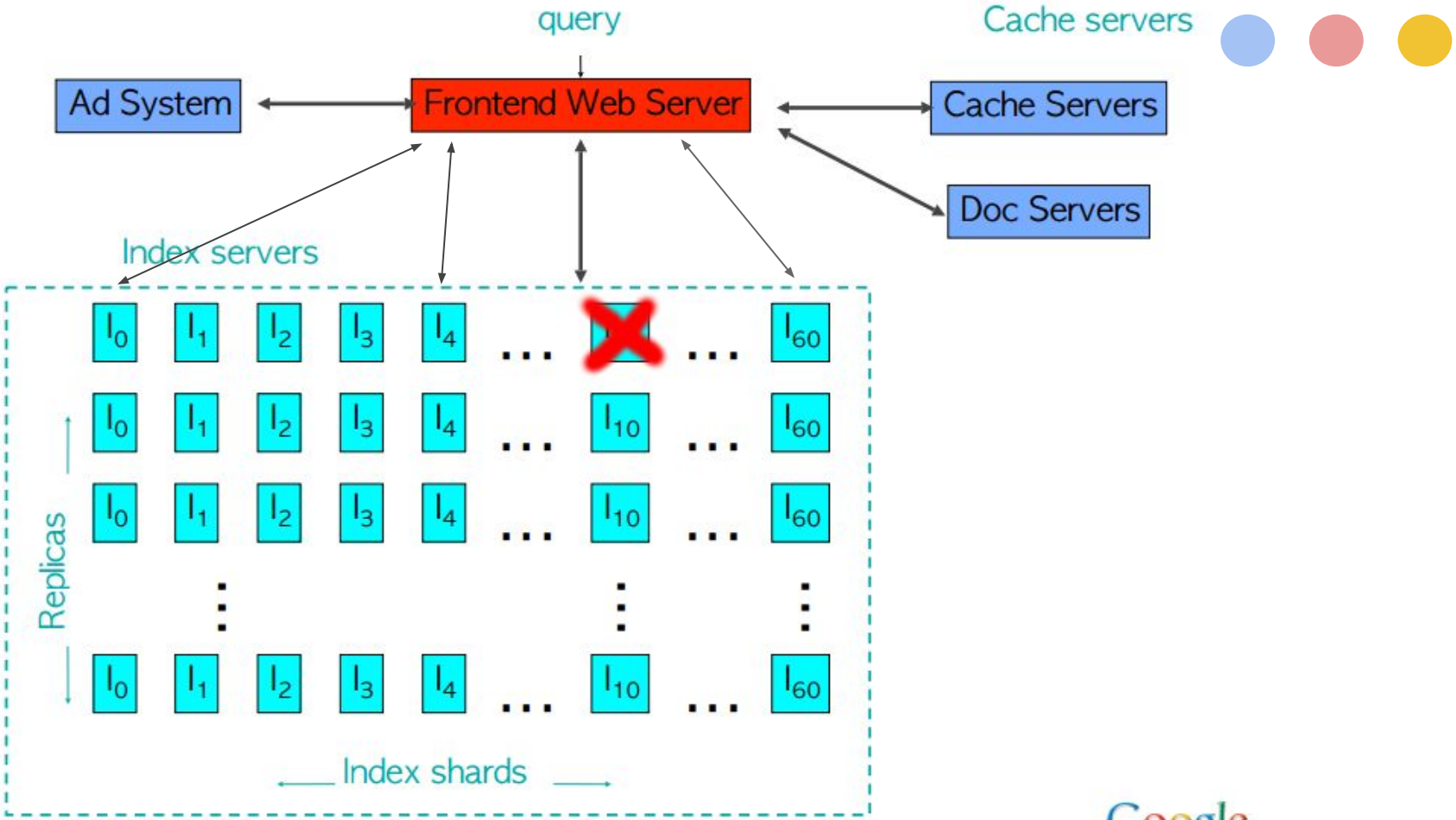
Inaccuracy for Resilience

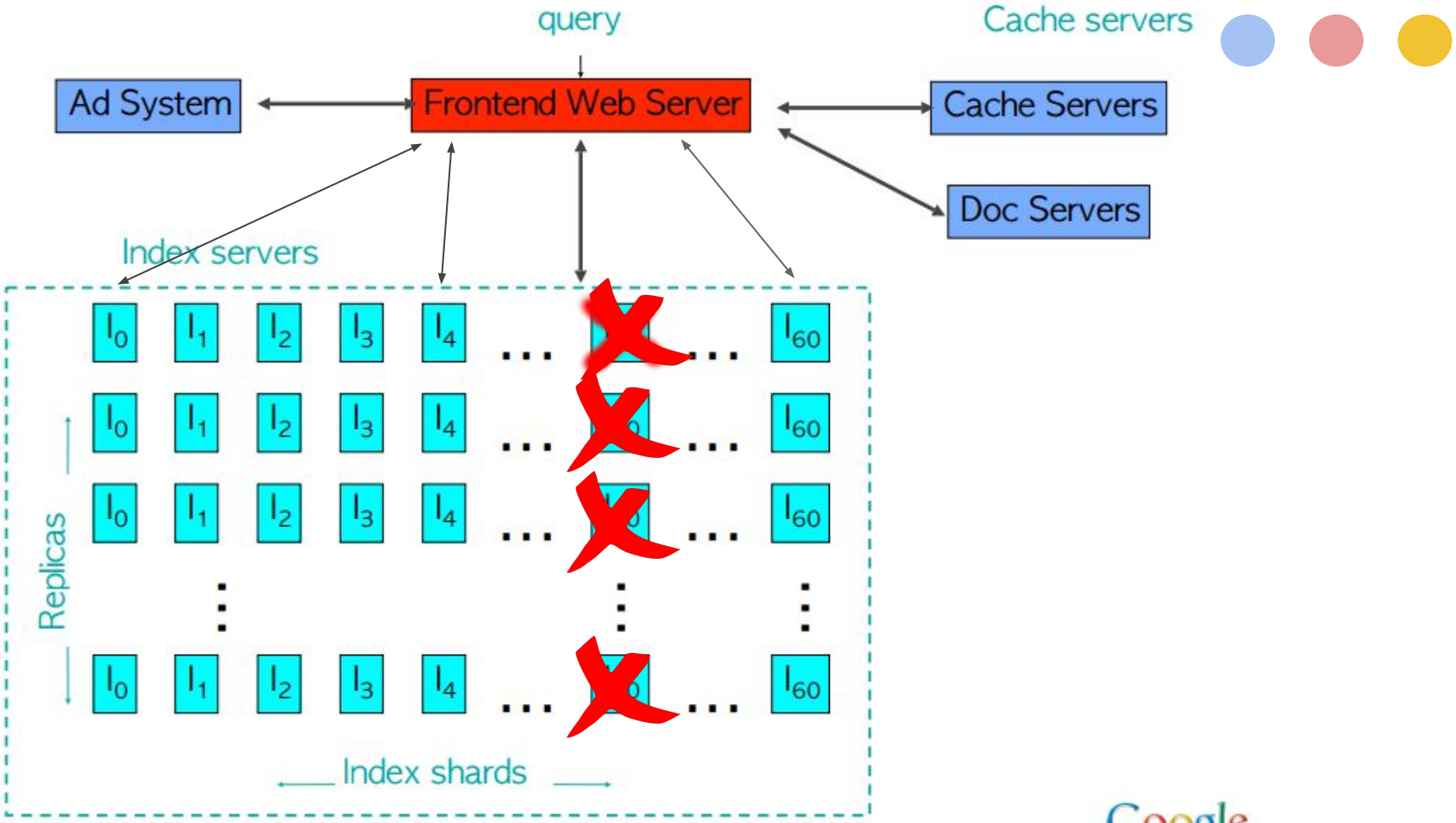
1. Task decomposition
2. Baseline for correctness
3. Criticality Testing
4. Distortion and timing models


query
Cache servers
Ad System
Frontend Web Server
Cache Servers
Doc Servers
Index servers
Replicas
Index shards


query
Cache servers
Ad System
Frontend Web Server
Cache Servers
Doc Servers
Index servers
I_0 I_1 I_2 I_3 I_4 ... I_{10} ... I_{60}
Replicas
Index shards

Google




query
Cache servers
Ad System
Frontend Web Server
Cache Servers
Doc Servers
Index servers
Replicas
Index shards

Robust & scaleable pipelines

Robust & scaleable pipelines

Leases for sharing & heartbeat

Robust & scaleable pipelines

Leases for sharing & heartbeat

Trade off inaccuracy for resilience & performance

Robust, scaleable pipelines
Leases for sharing & heartbeat
Tracing: Accurate for
resilient

CS research is timeless: use it to mitigate risk

Thanks

Ines Sombra

David Greenberg

Alex Hutcheson

Scott Zawalski

Karan Parikh

References

- T. Wurthinger, C. Wimmer et al. "One VM to Rule Them All"
- M. Rinard "Probabilistic Accuracy Bounds for Fault-Tolerant Computations that Discard Tasks"
- F. Corbato, M. Daggett, R. Daley "An Experimental Time-Sharing System"
- E. Dijkstra "Cooperating Sequential Processes"
- L. Lamport "Time, Clocks, and the Ordering of Events in a Distributed System"

References

- B. Oki, B. Liskov "Viewstamped Replication: A New Primary Copy Method to Support Highly-Available Distributed Systems"
- L. Lamport "The Part-Time Parliament"
- M. Welsh, D. Culler, E. Brewer "SEDA: An Architecture for Well-Conditioned, Scalable Internet Services"
- C. Gray, D. Cheriton "Leases: An Efficient Fault-Tolerant Mechanism for Distributed File Cache Consistency"
- S. Agarwal, B. Mozafari et al. "BlinkDB: Queries with Bounded Errors and Bounded Response Times on Very Large Data"

Should I read papers?

YES