# Distributed Common API for Measuring Performance

Alexander P. Sideropoulos

California Polytechnic State University, San Luis Obispo alexander@thequery.net

October 10, 2014

#### **Outline**

- Introduction
  - Motivation and Scope
  - Evaluation Criterion
  - Distributed CAMP
- 2 dCAMP Design
  - Architecture
  - Roles and Services
  - Operation
- dCAMP Details
  - ZeroMQ Protocols
- 4 dCAMP Analysis
  - Transparency
  - Scalability
- Evaluation and Conclusion

2 / 56

#### **Building Distributed Systems**

Due to the **plateauing speed** of individual processing units and encouraged by the **interconnectedness** of the internet at large, there exists a natural trend of distributing large, complex systems across multiple components locally and throughout the world.

#### **Building Distributed Systems**

In order to effectively build these systems, software practitioners must be able to test their system for **performance defects** as well as bottlenecks, all while the distributed system itself is responding to changes in availability and work load on its individual nodes.

### Distributed Performance Frameworks (DPF)

#### Distributed **Performance Testing** Frameworks:

- tools to evaluate performance of system
- both black box and white box
- instrumenting, collecting, analyzing, and visualizing distributed performance data

## Distributed Performance Frameworks (DPF)

#### Distributed **Performance Testing** Frameworks:

- tools to evaluate performance of system
- both black box and white box
- instrumenting, collecting, analyzing, and visualizing distributed performance data

#### Distributed **Performance Monitoring** Frameworks:

- considered part of the testing framework
- black box interface
- monitoring a distributed system or application
- mechanisms for triggering actions based on performance events

In order for practitioners and researchers alike to **effectively choose** a distributed performance framework, it is necessary to have a set criteria for evaluation. Presented here is an extended criterion of the general requirements presented by Zanikolas et al. for grid systems [1].

Data Delivery Models (original)

- Data Delivery Models (original)
- Security (original)

- Data Delivery Models (original)
- Security (original)
- Scalability (only consider good performance)

- Data Delivery Models (original)
- Security (original)
- Scalability (only consider good performance)
- Transparency (replaces Low Intrusiveness)

- Data Delivery Models (original)
- Security (original)
- Scalability (only consider good performance)
- Transparency (replaces Low Intrusiveness)
- Extensibility (removed)

- Data Delivery Models (original)
- Security (original)
- Scalability (only consider good performance)
- Transparency (replaces Low Intrusiveness)
- Extensibility (removed)
- Ompleteness (new)

- Data Delivery Models (original)
- Security (original)
- Scalability (only consider good performance)
- Transparency (replaces Low Intrusiveness)
- Extensibility (removed)
- Ompleteness (new)
- Validity (new)

- Data Delivery Models (original)
- Security (original)
- Scalability (only consider good performance)
- Transparency (replaces Low Intrusiveness)
- Extensibility (removed)
- Ompleteness (new)
- Validity (new)
- Portability (alternate)



### Distributed Common API for Measuring Performance

dCAMP is a distributed performance framework built on top of Mark Gabel and Michael Haungs' 2007 research on CAMP: a common API for measuring performance[2]. CAMP provides an accurate and "consistent method for retrieving system performance data from multiple platforms."

### Distributed Common API for Measuring Performance

dCAMP is a distributed performance framework built on top of Mark Gabel and Michael Haungs' 2007 research on CAMP: a common API for measuring performance[2]. CAMP provides an accurate and "consistent method for retrieving system performance data from multiple platforms."

*dCAMP* builds on this functionality and the authors' work validating *CAMP*'s accuracy by adding these core feature sets:

- stateful performance API
- distributed performance data aggregation
- performance filters and triggers
- simplistic fault tolerance

### System Architecture

*dCAMP* is designed as a **semi-centralized**, **hierarchical peer-to-peer system** utilizing the UNIX **Pipes and Filter** architectural pattern in which leaf nodes of the hierarchy collect data, filter out extraneous data, and send it up the pipe to a parent node which subsequently filters out more data and sends it up to another parent node.

The dCAMP distributed system is comprised of one or more nodes, each executing a role—a named grouping of a specific, known set of services. Each dCAMP service implements a specific function.

9 / 56

The dCAMP distributed system is comprised of one or more nodes, each executing a role—a named grouping of a specific, known set of services. Each dCAMP service implements a specific function.

• **Node**—rudimentary dCAMP functionality; topology communication, heartbeat monitoring, failure recovery

The dCAMP distributed system is comprised of one or more nodes, each executing a role—a named grouping of a specific, known set of services. Each dCAMP service implements a specific function.

- **Node**—rudimentary dCAMP functionality; topology communication, heartbeat monitoring, failure recovery
- **Sensor**—local metric gathering; essentially the dCAMP layer on top of OS/hardware APIs (via CAMP)

The dCAMP distributed system is comprised of one or more nodes, each executing a role—a named grouping of a specific, known set of services. Each dCAMP service implements a specific function.

- **Node**—rudimentary dCAMP functionality; topology communication, heartbeat monitoring, failure recovery
- **Sensor**—local metric gathering; essentially the dCAMP layer on top of OS/hardware APIs (via CAMP)
- Filter—metric filtering; throttling and thresholding

The *dCAMP* distributed system is comprised of one or more nodes, each executing a **role**—a named grouping of a specific, known set of **services**. Each *dCAMP* service implements a specific function.

- Node—rudimentary dCAMP functionality; topology communication, heartbeat monitoring, failure recovery
- Sensor—local metric gathering; essentially the dCAMP layer on top of OS/hardware APIs (via CAMP)
- Filter—metric filtering; throttling and thresholding
- Aggregation—metric aggregation; metric collection and calculation of multiple Sensor and/or Aggregation services

The *dCAMP* distributed system is comprised of one or more nodes, each executing a **role**—a named grouping of a specific, known set of **services**. Each *dCAMP* service implements a specific function.

- Node—rudimentary dCAMP functionality; topology communication, heartbeat monitoring, failure recovery
- Sensor—local metric gathering; essentially the dCAMP layer on top of OS/hardware APIs (via CAMP)
- Filter—metric filtering; throttling and thresholding
- Aggregation—metric aggregation; metric collection and calculation of multiple Sensor and/or Aggregation services
- Management-end-user control of dCAMP

The *dCAMP* distributed system is comprised of one or more nodes, each executing a **role**—a named grouping of a specific, known set of **services**. Each *dCAMP* service implements a specific function.

- Node—rudimentary dCAMP functionality; topology communication, heartbeat monitoring, failure recovery
- Sensor—local metric gathering; essentially the dCAMP layer on top of OS/hardware APIs (via CAMP)
- Filter—metric filtering; throttling and thresholding
- Aggregation—metric aggregation; metric collection and calculation of multiple Sensor and/or Aggregation services
- Management—end-user control of dCAMP
- Configuration—configuration replication; topology and configuration distribution



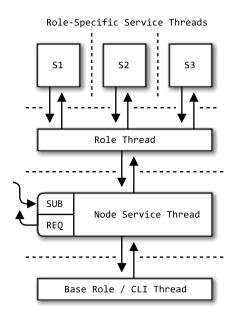
The dCAMP distributed system is comprised of one or more nodes, each executing a role—a named grouping of a specific, known set of services. Each dCAMP service implements a specific function.

A service's scope can vary depending on the node's level in the dCAMP topology.

- Root: master copy of the configuration, publishing new values as needed
- Collector: stores (and forwards) every update from the Root
- Metric: only stores updates relevant to the node

The dCAMP distributed system is comprised of one or more nodes, each executing a role—a named grouping of a specific, known set of services. Each dCAMP service implements a specific function.

Role	Service(s)
Root	Management, Aggregation, Filter, Configuration (Full)
Collector	Aggregation, Filter, Configuration (Full)
Metric	Sensor, Filter, Configuration (Partial)
Base	Node



The *Base* role must be running on each node for it to be part of the *dCAMP* distributed system. All other roles are launched from within the *Base* role.

• User promotes a *Root* node via the *dCAMP* CLI, specifying a configuration file and a Base node's address.

- User promotes a *Root* node via the *dCAMP* CLI, specifying a configuration file and a *Base* node's address.
- Root node connects to each Base node and begins the "discover" Topology Protocol.

- User promotes a *Root* node via the *dCAMP* CLI, specifying a configuration file and a *Base* node's address.
- Root node connects to each Base node and begins the "discover"
   Topology Protocol.
- Base nodes join the dCAMP system at any time, being assigned as Collector or Metric nodes in the topology.

- User promotes a *Root* node via the *dCAMP* CLI, specifying a configuration file and a *Base* node's address.
- Root node connects to each Base node and begins the "discover" Topology Protocol.
- Base nodes join the dCAMP system at any time, being assigned as Collector or Metric nodes in the topology.
- dCAMP runs in a steady state, nodes entering or exiting the system at any time.

- User promotes a *Root* node via the *dCAMP* CLI, specifying a configuration file and a Base node's address.
- Root node connects to each Base node and begins the "discover" Topology Protocol.
- **3** Base nodes join the dCAMP system at any time, being assigned as Collector or Metric nodes in the topology.
- dCAMP runs in a steady state, nodes entering or exiting the system at any time.
- User stops dCAMP by using the dCAMP CLI command.

- User promotes a *Root* node via the *dCAMP* CLI, specifying a configuration file and a *Base* node's address.
- Root node connects to each Base node and begins the "discover" Topology Protocol.
- Base nodes join the dCAMP system at any time, being assigned as Collector or Metric nodes in the topology.
- *dCAMP* runs in a steady state, nodes entering or exiting the system at any time.
- User stops dCAMP by using the dCAMP CLI command.
- 6 Root node begins the "stop" Topology Protocol.

- User promotes a *Root* node via the *dCAMP* CLI, specifying a configuration file and a *Base* node's address.
- Root node connects to each Base node and begins the "discover" Topology Protocol.
- Base nodes join the dCAMP system at any time, being assigned as Collector or Metric nodes in the topology.
- *dCAMP* runs in a steady state, nodes entering or exiting the system at any time.
- User stops dCAMP by using the dCAMP CLI command.
- 6 Root node begins the "stop" Topology Protocol.
- Collector and Metric nodes exit the topology and revert to Base nodes.



- User promotes a *Root* node via the *dCAMP* CLI, specifying a configuration file and a *Base* node's address.
- Root node connects to each Base node and begins the "discover" Topology Protocol.
- Base nodes join the dCAMP system at any time, being assigned as Collector or Metric nodes in the topology.
- *dCAMP* runs in a steady state, nodes entering or exiting the system at any time.
- User stops dCAMP by using the dCAMP CLI command.
- 6 Root node begins the "stop" Topology Protocol.
- Collector and Metric nodes exit the topology and revert to Base nodes.
- 8 Root node exits, reverting to Base node.



# Steady-State Operation

 Performance counters are sampled, filtered, reported, and logged by the Metric nodes at regular intervals according to the dCAMP Configuration.

# Steady-State Operation

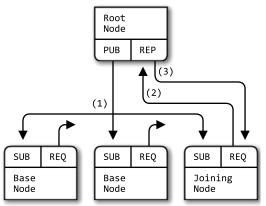
- Performance counters are sampled, filtered, reported, and logged by the Metric nodes at regular intervals according to the dCAMP Configuration.
- Performance counters received from child nodes are aggregated, filtered, reported, and logged by *Collector* nodes at regular intervals according to the dCAMP Configuration.

# Steady-State Operation

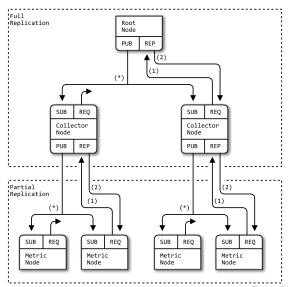
- Performance counters are sampled, filtered, reported, and logged by the Metric nodes at regular intervals according to the dCAMP Configuration.
- Performance counters received from child nodes are aggregated, filtered, reported, and logged by *Collector* nodes at regular intervals according to the dCAMP Configuration.
- Performance counters received from child nodes are aggregated and logged by Root node for later processing (e.g. graphing metrics during a test scenario or correlating statistics with a distributed event log).

#### **Topology Protocol**

The *dCAMP* distributed topology is dynamically established as the *Root* node sends out its discovery message and receives join messages from *Base* nodes. When a *Base* node responds to the *Root*, the *Base* node is given its assignment.



# Configuration Protocol



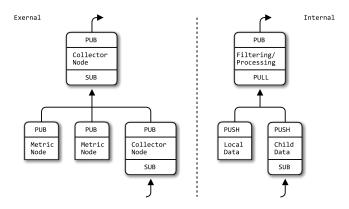
# Configuration Protocol

The dCAMP configuration replication algorithm generally adheres to the Clustered Hashmap Protocol[?].

```
config-replication = *update / snap-sync
                   = P-KVPUB / P-HUGZ
update
                   = C-ICANHAZ ( *P-KVSYNC P-KTHXBAI )
snap-sync
```

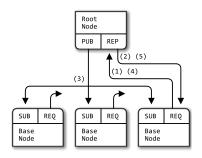
#### Data Protocol

There are two data flow protocols in the *dCAMP* system: the **external protocol** for data flowing between nodes and the **internal protocol** for data flowing between components within a single node.



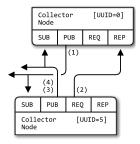
### Recovery Protocols: Promotion

The dCAMP Recovery Protocols are used for the Promotion and Election algorithms and use the same base messages as the Topology Protocol.



### Recovery Protocols: Election

The *dCAMP* Recovery Protocols are used for the Promotion and Election algorithms and use the same base messages as the Topology Protocol.



### Strategy

To measure the impact of dCAMP on a monitored process, a workload is defined and measured with and without dCAMP active. The measured difference in performance of the monitored process is defined to be dCAMP's monitoring overhead.

#### Workload

**Client:** Apache JMeter (v2.11) on a MacBook Pro (2.7GHz Core i7, 8GB 1333MHz DDR3, SSD) running OSX 10.9.

**Server:** default-configured Apache instance on a Lenovo Thinkpad (dual 2.16GHz Centrino Duo T2600, 2GB 667MHz DDR2, SATA) running Ubuntu 13.10.

Each test run includes 18 different load points, scaling the number of client threads from 2 to 2048. For every load point, the threads continuously (in this order):

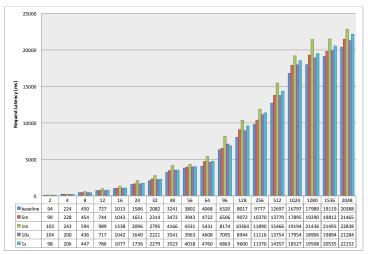
- 1 load a static home page,
- Oload a PHP page which calculates the 25th Fibonacci number, and
- 3 download a 5MB file of random binary data.

## Configuration

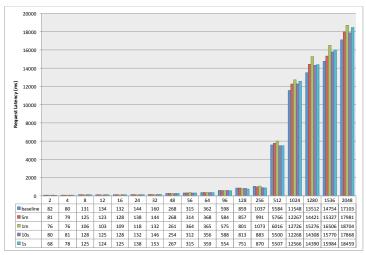
Each dCAMP configuration level monitors four global metrics and three process-specific metrics on the Apache process(es). The global metrics are CPU usage (proc), memory usage (mem), disk throughput (disk), and network throughput (net); the Apache metrics are CPU usage (apache\_cpu), memory usage (apache\_mem), and combined disk/network throughput (apache\_io).

- baseline dCAMP off
- 5m all metrics every 300 seconds, heartbeats every 60 seconds
- 1m − all metrics every 60 seconds, heartbeats every 60 seconds
- **10s** global metrics every 300 seconds, Apache metrics every 10 seconds, heartbeats every 300 seconds
- 1s global metrics every 300 seconds, Apache metrics every 1 second, heartbeats every 300 seconds

#### **CPU-Bound Results**



#### Disk-Bound Results



#### **Observations**

- When nodes are not expected to fail frequently, using longer heartbeat periods reduces the impact dCAMP has on the system.
- It is better to monitor a process using a faster sample period than an entire system using a slower sample period.
- The dCAMP system impact is noticeable but a considerably smaller factor than the impact hardware limitations have on performance monitoring.
- Holding all else constant, slower sample periods have an obviously lower impact on system performance compared to faster sample periods.
- Possibly using *dCAMP*'s reporting threshold, system impact can be minimized while still maintaining fine sample granularity.

## Strategy

One of the primary measures of scalability for a distributed system is its network traffic.[1] By simulating successively larger *dCAMP* systems (with respect to node count), one can extrapolate *dCAMP*'s effectiveness at monitoring large distributed systems and how to best configure its metric collections.

#### Workload

dCAMP is setup to monitor a machine's global metrics, scaling the number of **simulated nodes in the dCAMP system from 3 nodes up to 200 nodes**. The metric configuration is kept constant for each test run. As dCAMP starts, monitors in steady state, and shuts down, the machine's network traffic is monitored and recorded every five seconds.

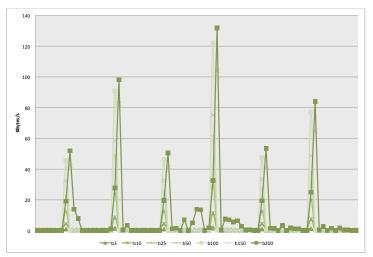
The test machine is a MacBook Pro (2.7GHz Core i7, 8GB 1333MHz DDR3, SSD) running OSX 10.9. All simulated *dCAMP* nodes use endpoints on the machine's loopback interface, and only the loopback interface traffic is monitored. The machine is otherwise entirely idle during the test runs.

### Configuration

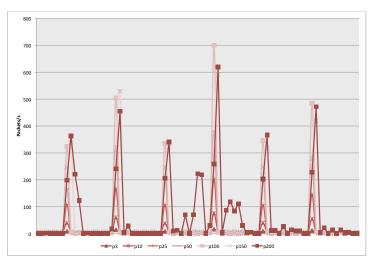
*dCAMP* is configured to monitor and report the below **global metrics**, using a heartbeat of 60 seconds.

- CPU usage every 60 seconds
- total disk throughput every 120 seconds
- total network throughput every 120 seconds
- memory usage every 60 seconds

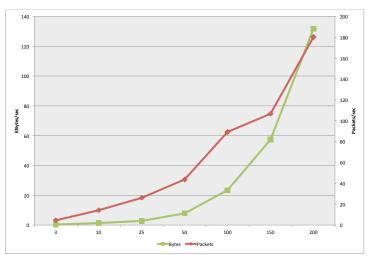
# Steady State Network Bytes



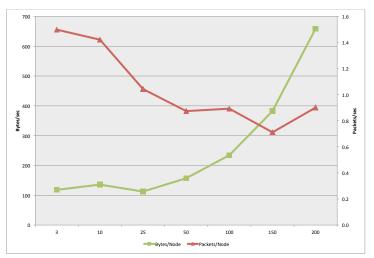
## Steady State Network Packets



## Average Network Utilization



## Average Network Utilization Per Node



#### **Observations**

- Sparklines of each load point show the same pattern: highest network traffic occurs during start up and then also on shutdown. This pattern follows the design of *dCAMP* which uses a chatty configuration protocol and a terse data protocol.
- As the node count increases, the rate at which bytes/packets are sent and received increases. This correlates with the larger configuration which dCAMP must track as well as the additional nodes sending and receiving data.
- Looking at the same values but also relating them to the number of nodes in the system, one sees the configuration size grows faster than the number of nodes.
- The number of messages being sent per node actually goes down and levels off just under 1 packet per node per second. This can be attributed to the fact that the number of Sensor nodes increases faster in relation to the number of *Collector* nodes.
- As this ratio increases, it is expected the number of messages per source

dcamp

contributions

#### References



[Zanikolas, 2005] Zanikolas,, Serafeim and Sakellariou,, Rizos (2005)

A taxonomy of grid monitoring systems

Future Gener. Comput. Syst. 21(1), 0167-739X, p163-188.



[Gabel, 2007] Gabel, Mark and Haungs, Michael (2007)

CAMP: a common API for measuring performance

LISA'07: Proceedings of the 21st conference on Large Installation System Administration Conference, p1–14

- 6 Appendix
  - Terminology
  - Evaluation Criterion Details
  - Related Work
  - ZeroMQ: Sockets and Patterns
  - ZeroMQ: Messaging Patterns
  - dCAMP Configuration
  - dCAMP Metrics

### **Terminology**

- dCAMP Service: Services are a way of logically grouping functions within the dCAMP system. Each service is implemented in dCAMP as an independent thread.
- dCAMP Role: Roles in the dCAMP system are groupings of one or more dCAMP services. There does not exist a one-to-one relationship between roles and services.
  - ZeroMQ: ZeroMQ is a message queuing framework which allows a developer to build distributed systems by focusing on the data and implementing simple design patterns.
- ZeroMQ Address: A ØMQ address is the combination of network host identifier (i.e. an IP Address or resolvable name) and Internet socket port number.
- ZeroMQ Endpoint: An endpoint is the combination of any ZeroMQ transport (pgm, inproc, ipc, or tcp) and a  $\emptyset$ MQ address.

- Performance Metric: Performance metrics are any data about a given node relating to its throughput, capacity, utilization, or latency.
- Metric Sampling: Metric collection or sampling is the process of measuring metrics on a given node.
- Metric Reporting: Metric reporting is the process of sending sampled metrics to a parent node.
- Metric Aggregation: Metric aggregation is the process of combining metrics from multiple nodes into a single metric, providing a coarser granularity for the performance metrics. Metrics are combined by calculating a sum, average, percent, or any other mathematically relevant operation.
- Metric Calculation: Metric calculation is the process of combining identical metrics from multiple timestamps into a single metric.

### Data Delivery Models

Monitoring information includes fairly static (e.g., software and hardware configuration of a given node) and dynamic events (e.g., current processor load, memory), which suggests the use of different measurement policies (e.g., periodic or on demand). In addition, consumer patterns may vary from sparse interactions to long lived subscriptions for receiving a constant stream of events. In this regard, the monitoring system must support both pull and push data delivery models.

## Security

Certain scenarios may require a monitoring service to support security services such as access control, single or mutual authentication of parties, and secure transport of monitoring information.

# Scalability

Monitoring systems have to cope efficiently with a growing number of resources, events and users. This scalability can be achieved as a result of good performance which guarantees that a monitoring system will achieve the needed throughput within an acceptable response time in a variety of load scenarios.

### **Transparency**

Transparency refers to the lack of impact a distributed performance framework makes on the system being monitored. As [1] states, it is "typically measured as a function of host (processor, memory, I/O) and network load (bandwidth) generated by the collection, processing and distribution of events." If a framework lacks transparency it will fail to allow the underlying distributed system to perform well and will produce inaccurate performance measurements, thereby reducing its Scalability and destroying its Validity.

#### Completeness

The Completeness of a distributed performance framework refers to the exhaustiveness to which it gathers performance metrics. At a minimum, a framework must provide interfaces for measuring and aggregating performance data about a system's processor, memory, disk, and network usage. Several distributed performance frameworks provide further detailed performance metrics about the given distributed system being monitored, but this is usually at the cost of Portability.

## **Validity**

A distributed performance framework is only as good as the data is produces; if the sensors or gathering techniques are inaccurate, then the data is useless at best, misleading at worst. Validity of a framework is achieved when the authors of a framework provide formal verification of its accuracy.

## **Portability**

A framework's ability to run on a completely heterogeneous distributed system without special considerations by the practitioner is what this work defines as Portability. More specifically, a portable framework has a unified API regardless of the system architecture, does not restrict itself to applications written in specific programming languages, and does not require practitioners to manually instrument their application code. This black box characteristic is vital for a viable distributed performance framework's effectiveness as it allows practitioners to focus on the performance data and not on a myriad of APIs for various architectures or languages.

related work

details

Publish/Subscribe

 ${\sf Request}/{\sf Reply}$ 

Pipeline

**Exclusive Pair** 

**Node Specification** 

Sample Specification

overview, dcamp metric groups, t3.1, pitfall