McCAT: Multi-cloud Cost-aware Transport

Sebastian Frischbier* TU Darmstadt

Alessandro Margara University of Lugano (USI)

Tobias Freudenreich† TU Darmstadt

Patrick Eugster Purdue University

David Eyers University of Otago

Peter Pietzuch Imperial College London

Motivation and idea. Distributed applications in use today, such as monitoring systems (e.g. Ganglia), data-parallel processing frameworks (e.g. Hadoop) and replicated databases (e.g. MySQL), were originally developed for use in clusters. Cluster applications often assume that the network is well-provisioned and its usage is effectively free, as long as it remains within capacity bounds.

As cluster applications are deployed in cloud environments, or even across multiple data centre sites, these assumptions no longer hold: cloud tenants would like to control the network usage of distributed cluster applications, e.g. to minimise their costs.

Our observation is that the network usage of many cluster applications can be reduced in a controlled fashion. For example, in a multi-site Ganglia deployment, it is possible to reduce network traffic between sites by aggregating multiple updates of monitored metrics before transmission, or by filtering metrics that are not being displayed to users. Similarly, a master-slave deployment of MySQL can multicast the replication stream to multiple sites, or filter the stream between sites to prioritise the transmission of updates depending on the query workload.

We propose to control the network usage of cluster applications by creating a **cost-aware transport-layer service**. Such a service can transparently (i) *filter* the data transmitted if it is ultimately not used by the application; (ii) *aggregate* multiple data items into one to save bandwidth by reducing precision; and (iii) *multicast* data items to avoid redundant unicast transmissions of the same data across sites.

A cost-aware transport service can apply these techniques to reduce network usage in order to stay within cost limits. In contrast to existing transport layer protocols, such a service must respect application semantics to reason about the impact of filter, aggregation and multicast operations on the application.

Introducing McCAT. To demonstrate the benefit and feasibility of this approach, we describe **McCAT**—a cost-aware many-to-many transport service based on [1] that offers tenants control over how cluster applications use the network. McCAT provides custom interceptors for TCP connections between nodes of a cluster application. After interception, it represents semantically meaningful data items as *messages*.

Messages are transported by a multi-site overlay network of McCAT brokers that support: (i) a *subscription mechanism* that uses a publish/subscribe communication model to filter messages; (ii) an *on-path aggregation* mechanism to combine multiple messages; and (iii) *application-layer multicast* to distribute messages efficiently across multiple sites.

McCAT applies the above mechanisms to ensure that the network usage between brokers stays within limits defined by the tenant. For example, tenants may specify a maximum bandwidth usage between brokers located at different sites. McCAT therefore exposes a trade-off between the costs of network usage in a cloud environment, and the quality achieved by cluster applications in terms of data completeness or precision.

Key features of McCAT are in more detail:

(i) Subscription mechanism: McCAT realises an early-filtering strategy, even for cluster applications that do not support this. McCAT decides if messages can be filtered without transmission based on the current interest of data consumers, which are expressed using subscriptions. For example, it becomes possible to discard metrics produced by Ganglia's `gmond` monitors if no `gmetad` nodes process these metrics.

In addition, McCAT can feed knowledge about consumers' interests back to the data producers, enabling them to avoid data production. For example, `gmond` monitors can be reconfigured to produce only metrics that are currently consumed by `gmetad` nodes. This reconfiguration relies on lightweight wrappers that McCAT provides for Ganglia daemons.

(ii) On-path aggregation: McCAT can reduce the number and size of transported messages by aggregating them in an application-specific manner: producers inform McCAT about acceptable aggregations for their data. For example, McCAT can (a) average multiple monitored values in Ganglia and send them as a single message; or (b) conserve messages by only sending messages when values change. For both mechanisms, McCAT permits cloud tenants to specify the maximum acceptable imprecision of their data.

(iii) Multicast communication: Since messages are routed via a multi-broker overlay network, McCAT can multicast messages to multiple data consumers in any part of the network topology. This means that it adds application-layer multicast support to applications that otherwise use unicast TCP connections. For example, McCAT can increase the network efficiency of MySQL's master-slave replication mechanism by multicasting replication messages to multiple slaves.

References

[1] S. Frischbier, A. Margara, T. Freudenreich, P. Eugster, D. Eyers, and P. Pietzuch. Aggregation for implicit invocations. In *AOSD'13*, Mar. 2013.

*Presenter, student

†Student

(Image by courtesy of Silvia Dacosta)

TECHNISCHE UNIVERSITÄT DARMSTADT · Università della Svizzera italiana · PURDUE UNIVERSITY · UNIVERSITY of OTAGO NEW ZEALAND · Imperial College London

McCAT: Multi-cloud Cost-Aware Transport

S. Frischbier, A. Margara, T. Freudenreich, P. Eugster, D. Eyers, P. Pietzuch

Reduce cost of running cluster applications in cloud environments

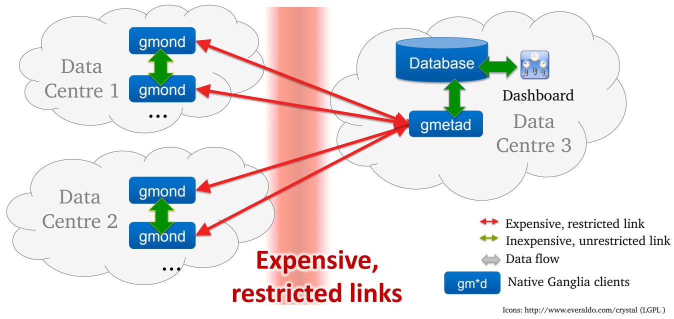
Cluster applications (e.g. Ganglia, MySQL, Hadoop) are **unaware of network usage** and assume:

- **Inexpensive network connections within one site**
- Well-provisioned network, effectively free usage

Leads to **high network costs** when running in public clouds, particularly across **multiple data centre sites**

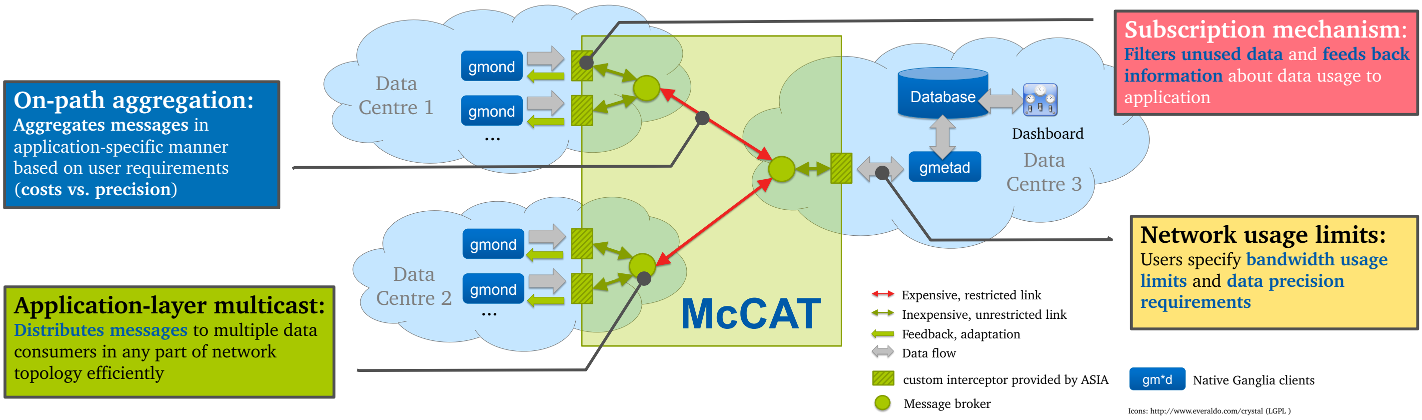
- **Wide-area network (WAN) links more expensive**
- Heterogeneous bandwidth, no multicast support

Example: **Ganglia** for distributed monitoring of cloud applications[2]



Trade off network usage for application benefit

McCAT cost-aware transport layer ensures that **network usage** between nodes stays within **tenant-specified limits.** **Uses** overlay network with broker nodes **with multiple techniques [1]**



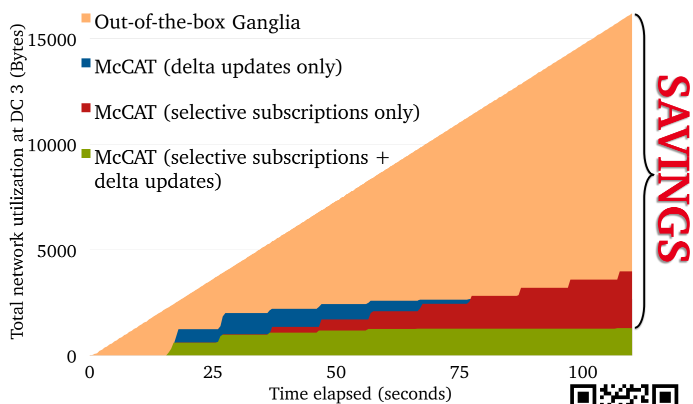
Running Ganglia using McCAT saves network costs

Deployment with **interceptors** for Ganglia **to cut WAN costs**

Experimental set-up: 3 data centres on Amazon EC2

Results: **Aggregation + Subscription reduces WAN costs**

- **Aggregation (delta updates) significantly reduces network usage**
- **Savings proportional to filter selectivity**



Ongoing work

- Apply McCAT to **MySQL master-slave replication**
- **Extend user control** over data imprecision

References:
1. S. Frischbier, A. Margara, T. Freudenreich, P. Eugster, D. Eyers, and P. Pietzuch. Aggregation for Implicit Invocations. In AOSD'13, Mar. 2013.
2. http://ganglia.sourceforge.net/

Primary contact:
Sebastian Frischbier, frischbier@dvs.tu-darmstadt.de
Website: http://www.dvs.tu-darmstadt.de/research/events/asia/