Improving Scalability and Fault Tolerance in an Application Management Infrastructure

Nikolay Topilski, Jeannie Albrecht,
and Amin Vahdat

Williams College & UC San Diego

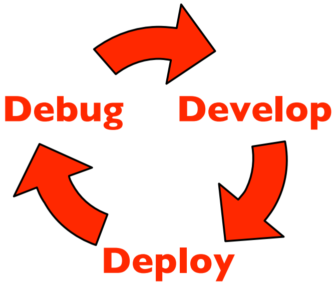
Large-Scale Computing

- Large-scale computing has many advantages
  - Increased computing power leads to improved performance, scalability, and fault tolerance

- Also introduces many new challenges
  - Building and managing distributed applications to leverage full potential of large-scale environments is difficult

Distributed Application Management

- Develop-Deploy-Debug cycle
  - Develop software
  - Deploy on distributed machines
  - Debug code when problems arise



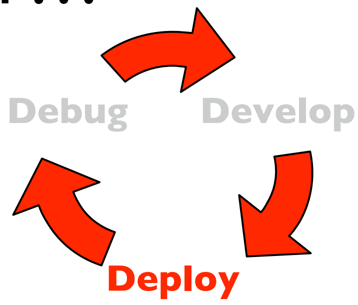
Debug
Develop
Deploy

- Management challenges in large-scale environments
  - Configuring resources
  - Detecting and recovering from failures
  - Achieving scalability and fault tolerance

- Research goal: Build an application management infrastructure that addresses these challenges

Deploying an Application

- Steps required to deploy an application
  1. Connect to each resource
  2. Download software
  3. Install software
  4. Run application
  5. Check for errors on each machine
  6. When we find an error, we start all over…
- A better alternative: Plush



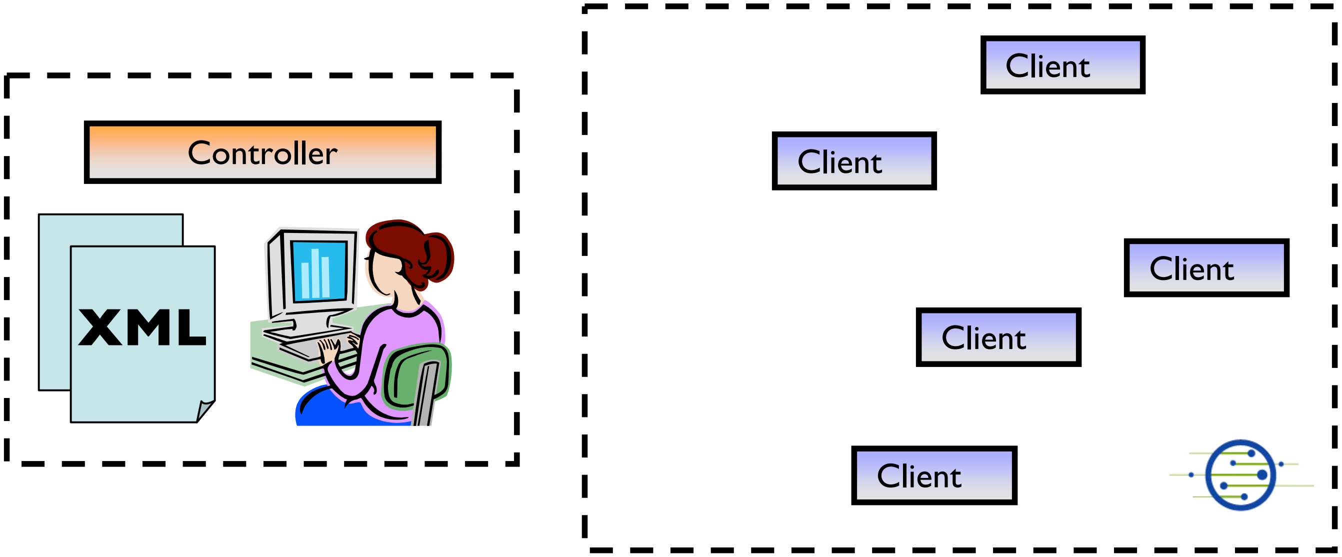
Plush

- Distributed application management infrastructure
  - Designed to simplify management of distributed applications
  - Help software developers cope with the challenges of large-scale computing
  - Support most applications in most environments

- Talk overview
  - Give brief overview of Plush architecture
  - Discuss scalability and fault tolerance limitations in original design
  - Investigate ways to improve limitations

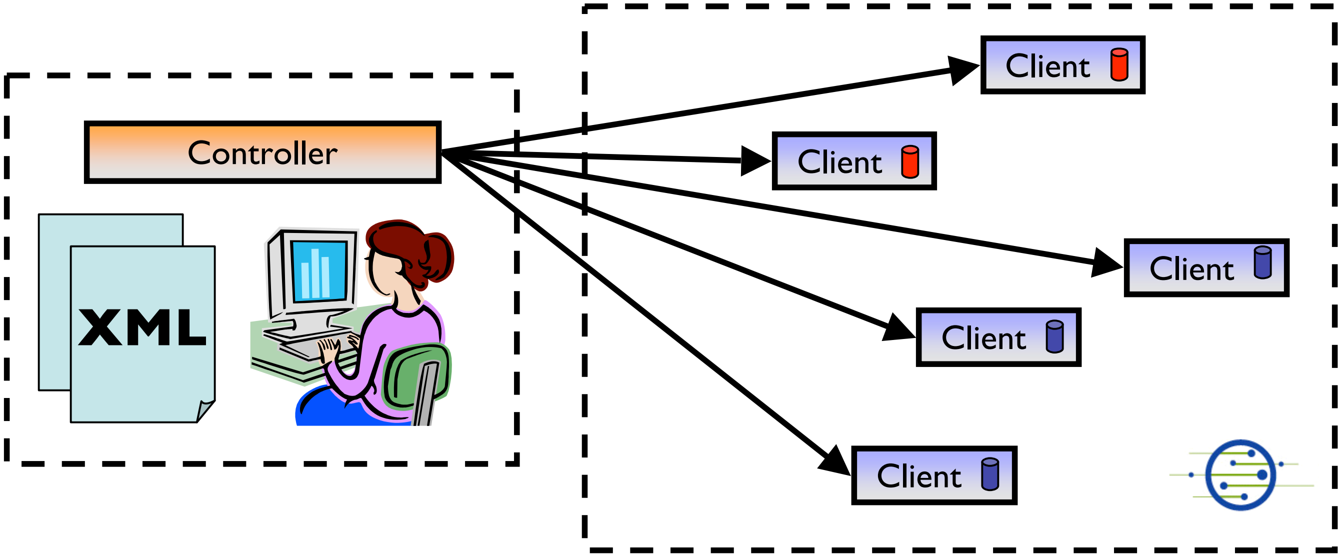
Plush Overview

- Plush consists of two main components:
  - Controller - runs on user’s Desktop
  - Client - runs on distributed resources
- To start application, user provides controller with application specification and resource directory (XML)



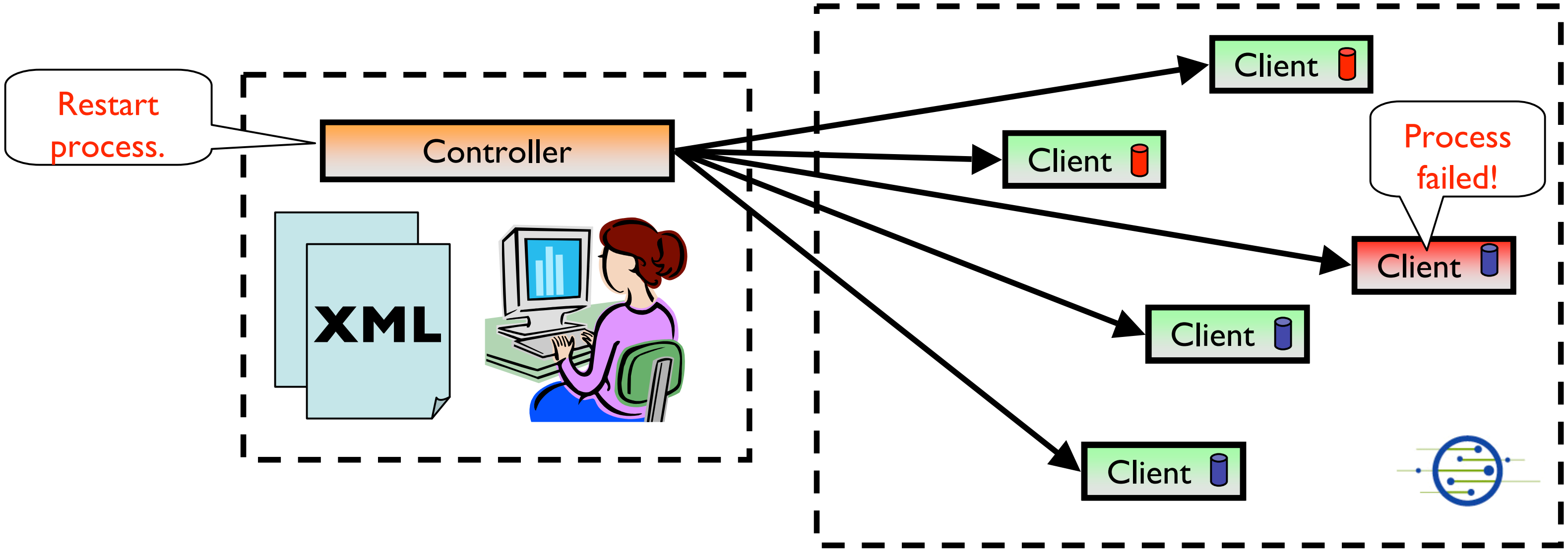
Plush Overview

- Controller makes direct TCP connection to each client process running remotely
  - Communication mesh forms star topology
- Controller instructs clients to download and install software (described in app spec)



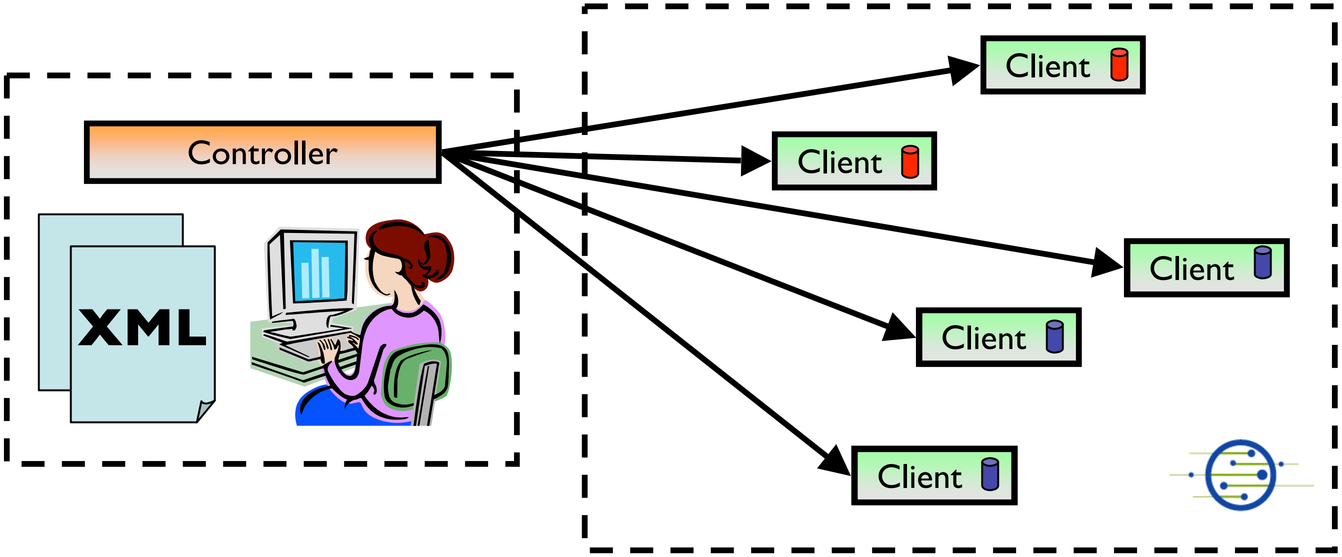
Plush Overview

- When all resources have been configured, controller instructs clients to begin execution
- Clients monitor processes for errors
  - Notify controller if failure occurs



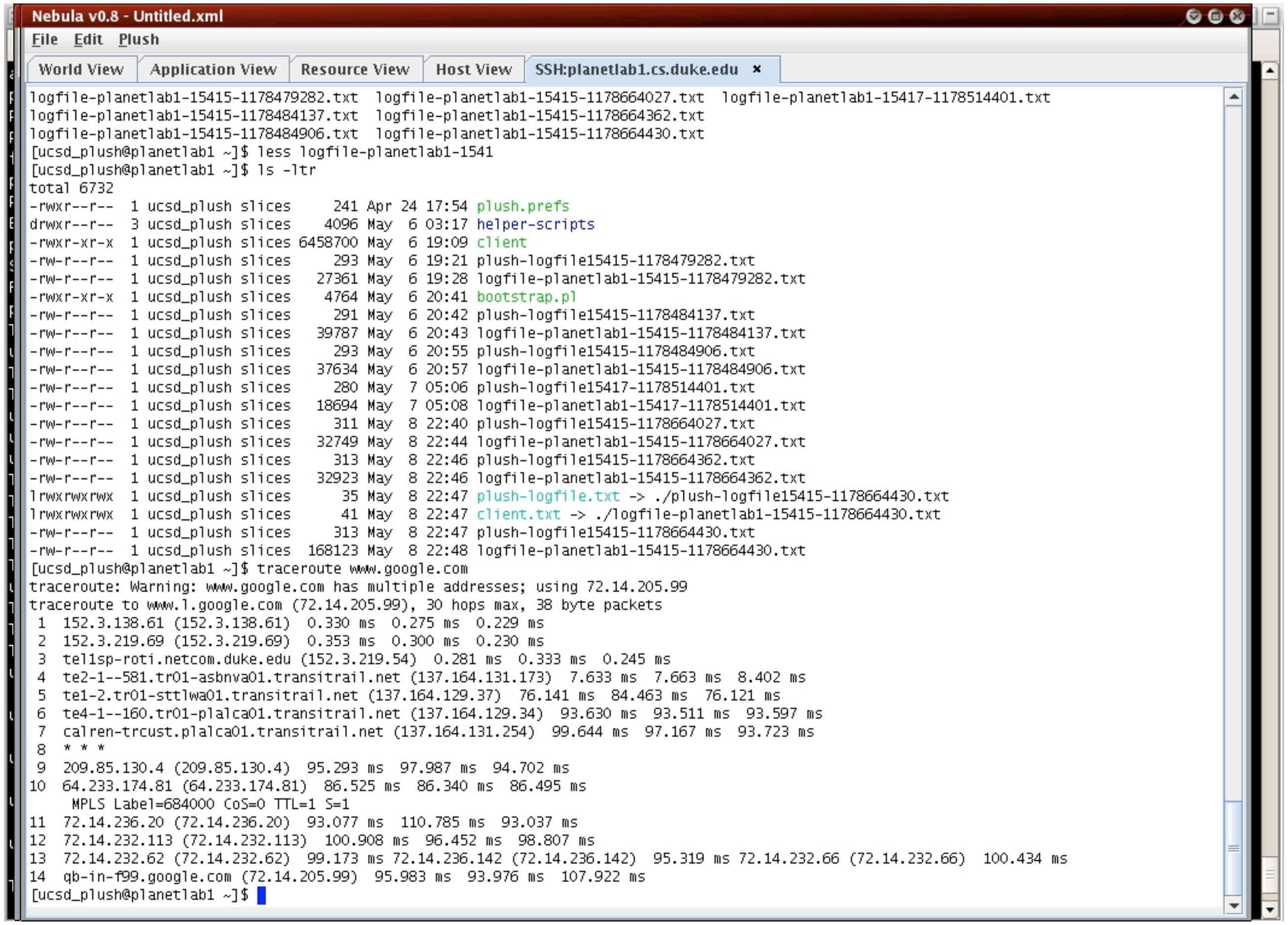
Plush Overview

- Once execution completes, controller instructs clients to "clean up"
  - Stop any remaining processes
  - Remove log files
  - Disconnect TCP connections



Plush User Interfaces

- Command-line interface used to interact with applications
- Nebula (GUI) allows users to describe, run, & visualize applications
- XML-RPC interface for managing applications programmatically



```
logfile-planetlab1-15415-1178479282.txt  logfile-planetlab1-15415-1178664027.txt  logfile-planetlab1-15417-1178514401.txt
logfile-planetlab1-15415-1178484137.txt  logfile-planetlab1-15415-1178664362.txt
logfile-planetlab1-15415-1178484906.txt  logfile-planetlab1-15415-1178664430.txt
[ucsd_plush@planetlab1 ~]$ less logfile-planetlab1-1541
[ucsd_plush@planetlab1 ~]$ ls -ltr
total 6732
-rwxr--r--  1 ucsd_plush slices     241 Apr 24 17:54 plush.prefs
drwxr--r--  3 ucsd_plush slices    4096 May  6 03:17 helper-scripts
-rwxr-xr-x  1 ucsd_plush slices 6458700 May  6 19:09 client
-rw-r--r--  1 ucsd_plush slices     293 May  6 19:21 plush-logfile15415-1178479282.txt
-rw-r--r--  1 ucsd_plush slices   27361 May  6 19:28 logfile-planetlab1-15415-1178479282.txt
-rwxr-xr-x  1 ucsd_plush slices    4764 May  6 20:41 bootstrap.pl
-rw-r--r--  1 ucsd_plush slices     291 May  6 20:42 plush-logfile15415-1178484137.txt
-rw-r--r--  1 ucsd_plush slices   39787 May  6 20:43 logfile-planetlab1-15415-1178484137.txt
-rw-r--r--  1 ucsd_plush slices     293 May  6 20:55 plush-logfile15415-1178484906.txt
-rw-r--r--  1 ucsd_plush slices   37634 May  6 20:57 logfile-planetlab1-15415-1178484906.txt
-rw-r--r--  1 ucsd_plush slices     280 May  7 05:06 plush-logfile15417-1178514401.txt
-rw-r--r--  1 ucsd_plush slices   18694 May  7 05:08 logfile-planetlab1-15417-1178514401.txt
-rw-r--r--  1 ucsd_plush slices     311 May  8 22:40 plush-logfile15415-1178664027.txt
-rw-r--r--  1 ucsd_plush slices   32749 May  8 22:44 logfile-planetlab1-15415-1178664027.txt
-rw-r--r--  1 ucsd_plush slices     313 May  8 22:46 plush-logfile15415-1178664362.txt
-rw-r--r--  1 ucsd_plush slices   32923 May  8 22:46 logfile-planetlab1-15415-1178664362.txt
lrwxrwxrwx  1 ucsd_plush slices      35 May  8 22:47 plush-logfile.txt -> ./plush-logfile15415-1178664430.txt
lrwxrwxrwx  1 ucsd_plush slices      41 May  8 22:47 client.txt -> ./logfile-planetlab1-15415-1178664430.txt
-rw-r--r--  1 ucsd_plush slices     313 May  8 22:47 plush-logfile15415-1178664430.txt
-rw-r--r--  1 ucsd_plush slices  168123 May  8 22:48 logfile-planetlab1-15415-1178664430.txt
[ucsd_plush@planetlab1 ~]$ traceroute www.google.com
traceroute: Warning: www.google.com has multiple addresses; using 72.14.205.99
traceroute to www.l.google.com (72.14.205.99), 30 hops max, 38 byte packets
 1  152.3.138.61 (152.3.138.61)  0.330 ms  0.275 ms  0.229 ms
 2  152.3.219.69 (152.3.219.69)  0.353 ms  0.300 ms  0.230 ms
 3  tel1sp-roti.netcom.duke.edu (152.3.219.54)  0.281 ms  0.333 ms  0.245 ms
 4  te2-1--581.tr01-asbnva01.transitrail.net (137.164.131.173)  7.633 ms  7.663 ms  8.402 ms
 5  te1-2.tr01-sttlwa01.transitrail.net (137.164.129.37)  76.141 ms  84.463 ms  76.121 ms
 6  te4-1--160.tr01-plalca01.transitrail.net (137.164.129.34)  93.630 ms  93.511 ms  93.597 ms
 7  calren-trcust.plalca01.transitrail.net (137.164.131.254)  99.644 ms  97.167 ms  93.723 ms
 8  * * *
 9  209.85.130.4 (209.85.130.4)  95.293 ms  97.987 ms  94.702 ms
10  64.233.174.81 (64.233.174.81)  86.525 ms  86.340 ms  86.495 ms
     MPLS Label=684000 CoS=0 TTL=1 S=1
11  72.14.236.20 (72.14.236.20)  93.077 ms  110.785 ms  93.037 ms
12  72.14.232.113 (72.14.232.113)  100.908 ms  96.452 ms  98.807 ms
13  72.14.232.62 (72.14.232.62)  99.173 ms 72.14.236.142 (72.14.236.142)  95.319 ms 72.14.232.66 (72.14.232.66)  100.434 ms
14  qb-in-f99.google.com (72.14.205.99)  95.983 ms  93.976 ms  107.922 ms
[ucsd_plush@planetlab1 ~]$ 
```

Limitations

- Plush was designed with PlanetLab in mind…
  - … in 2004!
  - PlanetLab grew from 300 machines to 800+
- Plush now supports execution in a variety of environments in addition to PlanetLab
  - Some have 1000+ resources

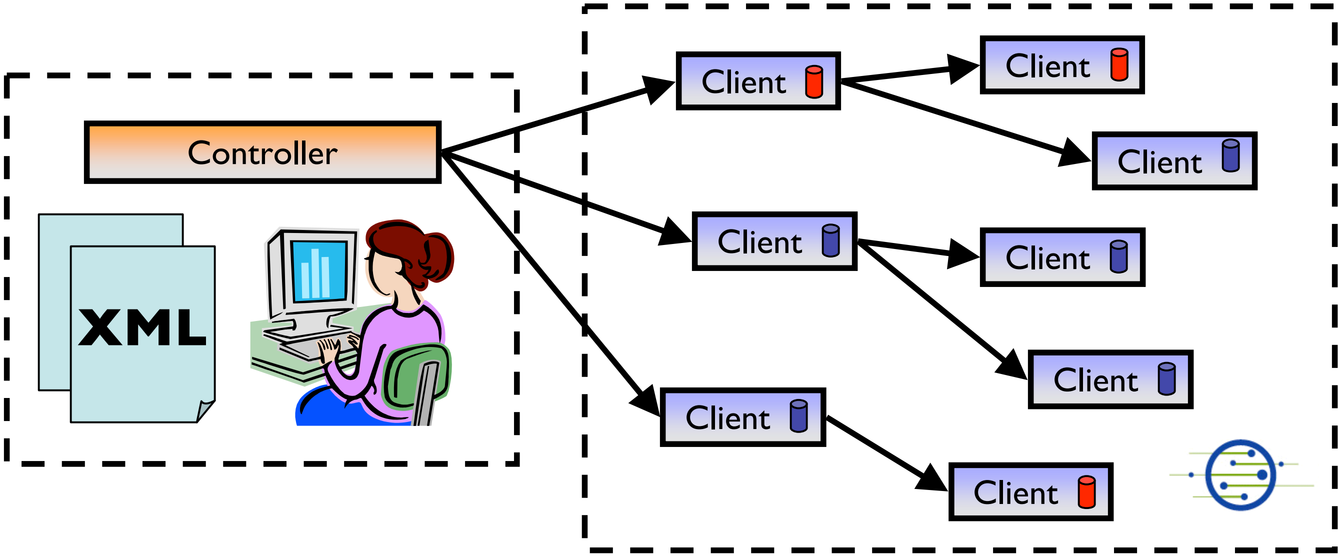
- Problems
  - Star topology does not scale beyond ~300 resources
  - Tree topology scales but is not resilient to failure

Insights

- We need a *resilient* overlay tree in place of the star
- Lots of people have already studied overlay tree building algorithms
- Mace is a framework for building overlays
  - Developed at UCSD
  - Simplifies development through code reuse
- Solution: Combine Plush with overlay tree provided by Mace!
  - Allow us to explore different tree building protocols
  - Leverage existing research in overlay networks without “reinventing the wheel”
  - Improve scalability and fault tolerance of Plush

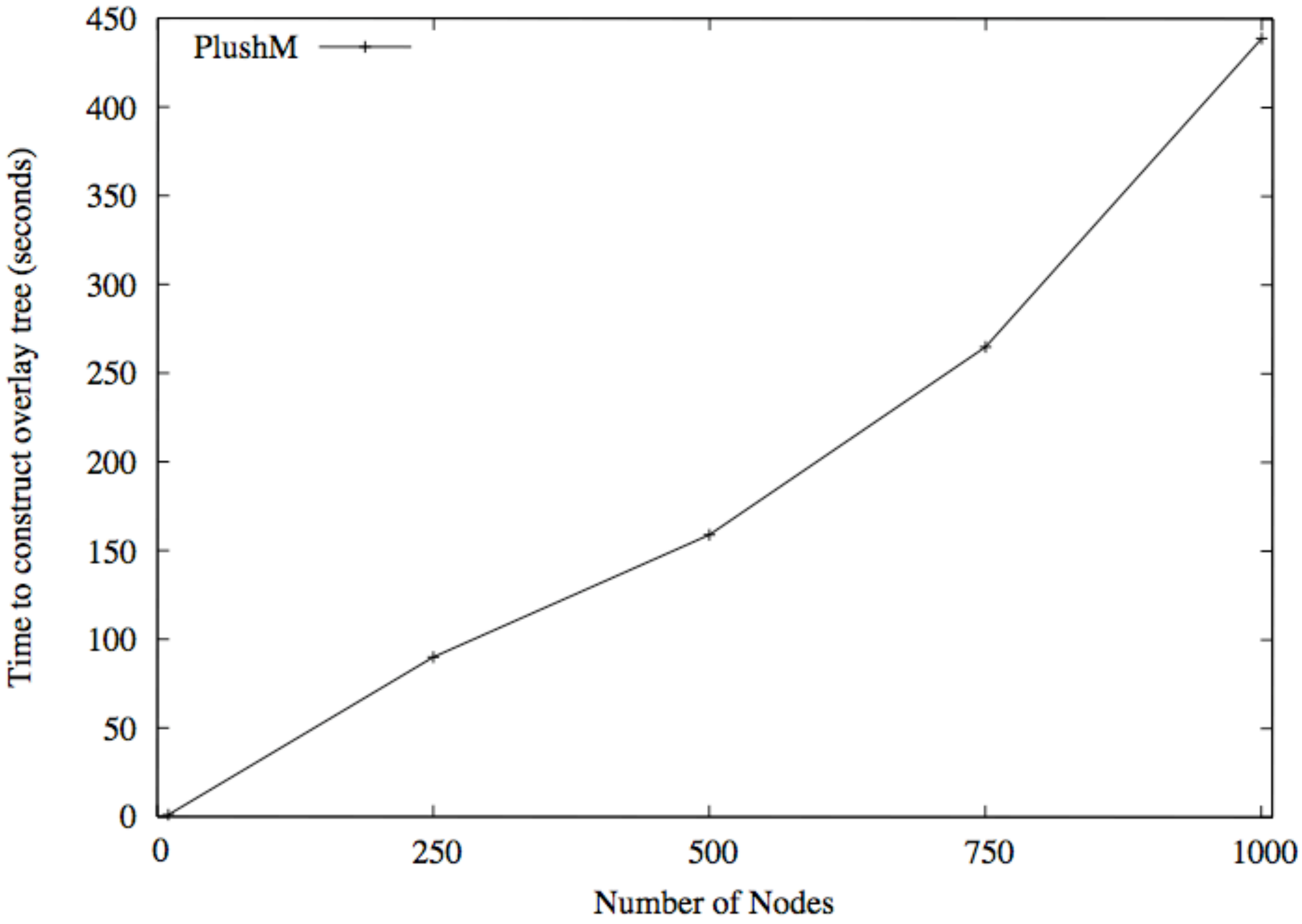
Introducing PlushM

- We extended the existing communication fabric in Plush to allow interaction with Mace (⇒ PlushM)
- PlushM still uses same abstractions for application management as Plush
- We chose RandTree as our initial overlay topology
  - Random overlay tree that reconfigures when failure occurs



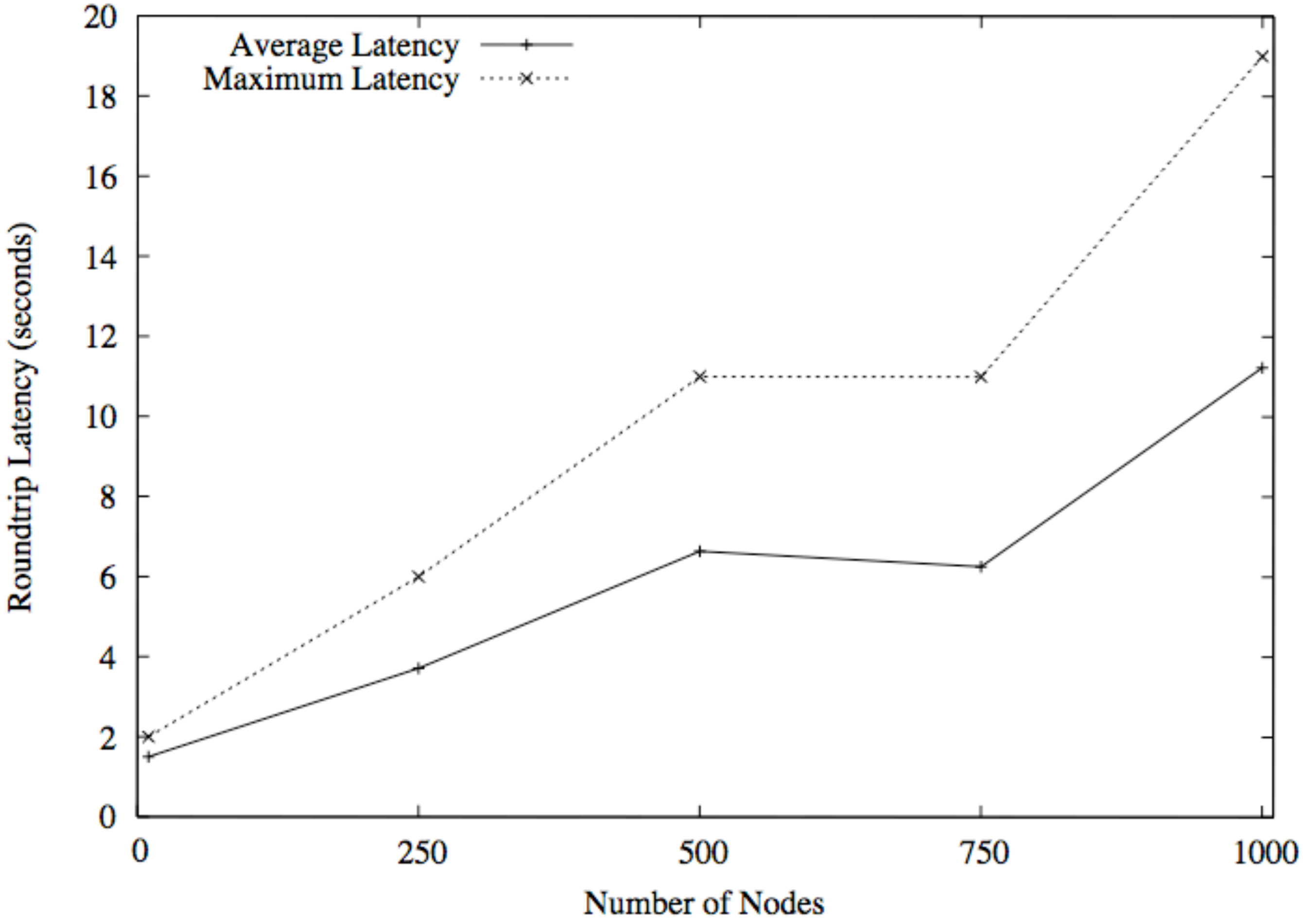
Evaluating Scalability

- Overlay tree construction time



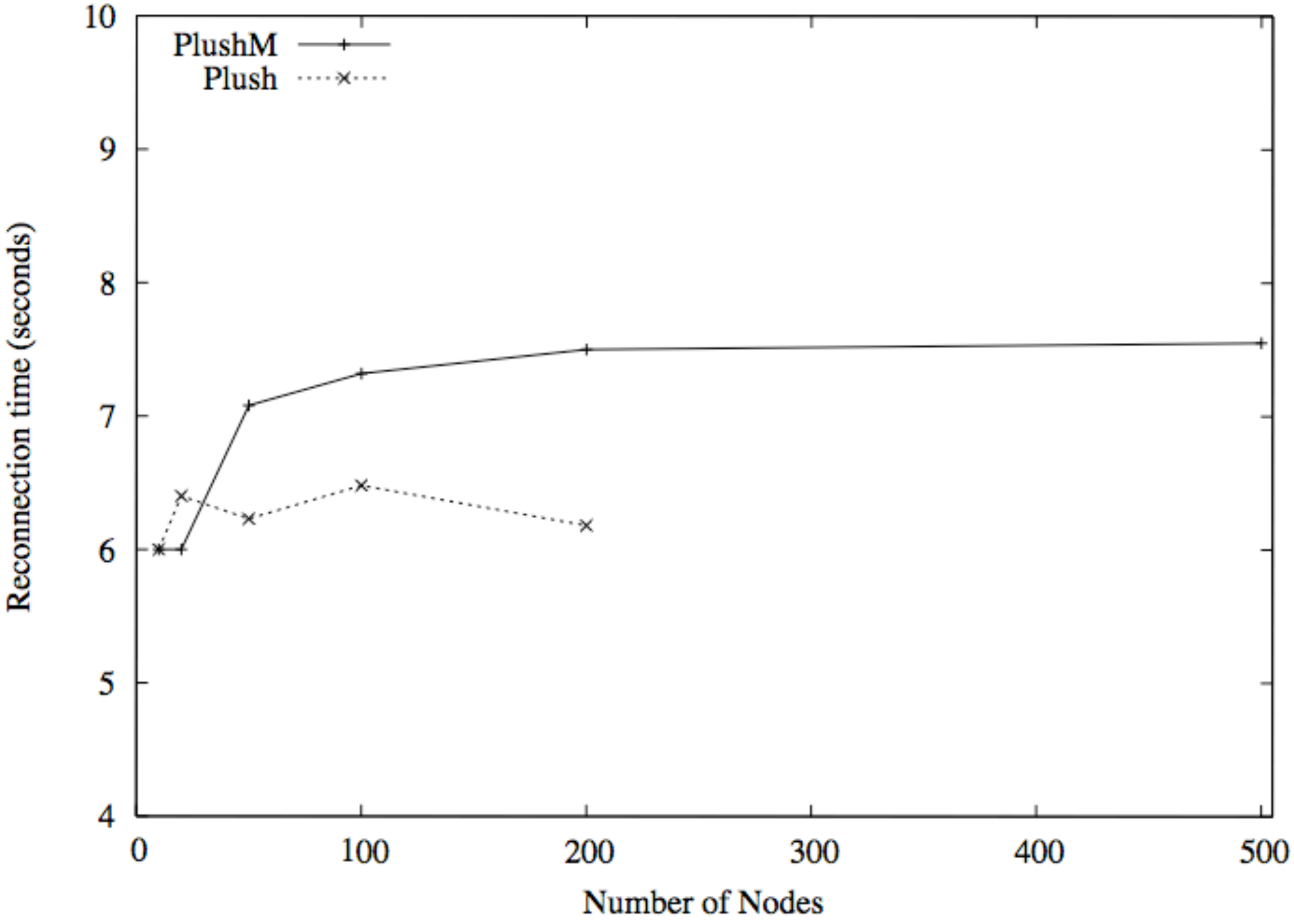
Evaluating Scalability

- Message propagation time



Evaluating Fault Tolerance

- Reconfiguration time after disconnect (ModelNet)



Conclusions and Future Work

- Plush provides distributed application management in a variety of environments
  - Original design has scalability/fault tolerance limitations in large-scale clusters
- PlushM replaces Plush's communication infrastructure with Mace overlay to provide better scalability (1000 resources) and fault tolerance
- Future work
  - Evaluate PlushM on larger topologies
  - Investigate the user of other Mace overlays in addition to RandTree
  - Explore ways to improve PlushM performance

Thank you!

Plush http://plush.cs.williams.edu

Mace http://mace.ucsd.edu

Email

ntopilsk@cs.ucsd.edu

jeannie@cs.williams.edu

vahdat@cs.ucsd.edu