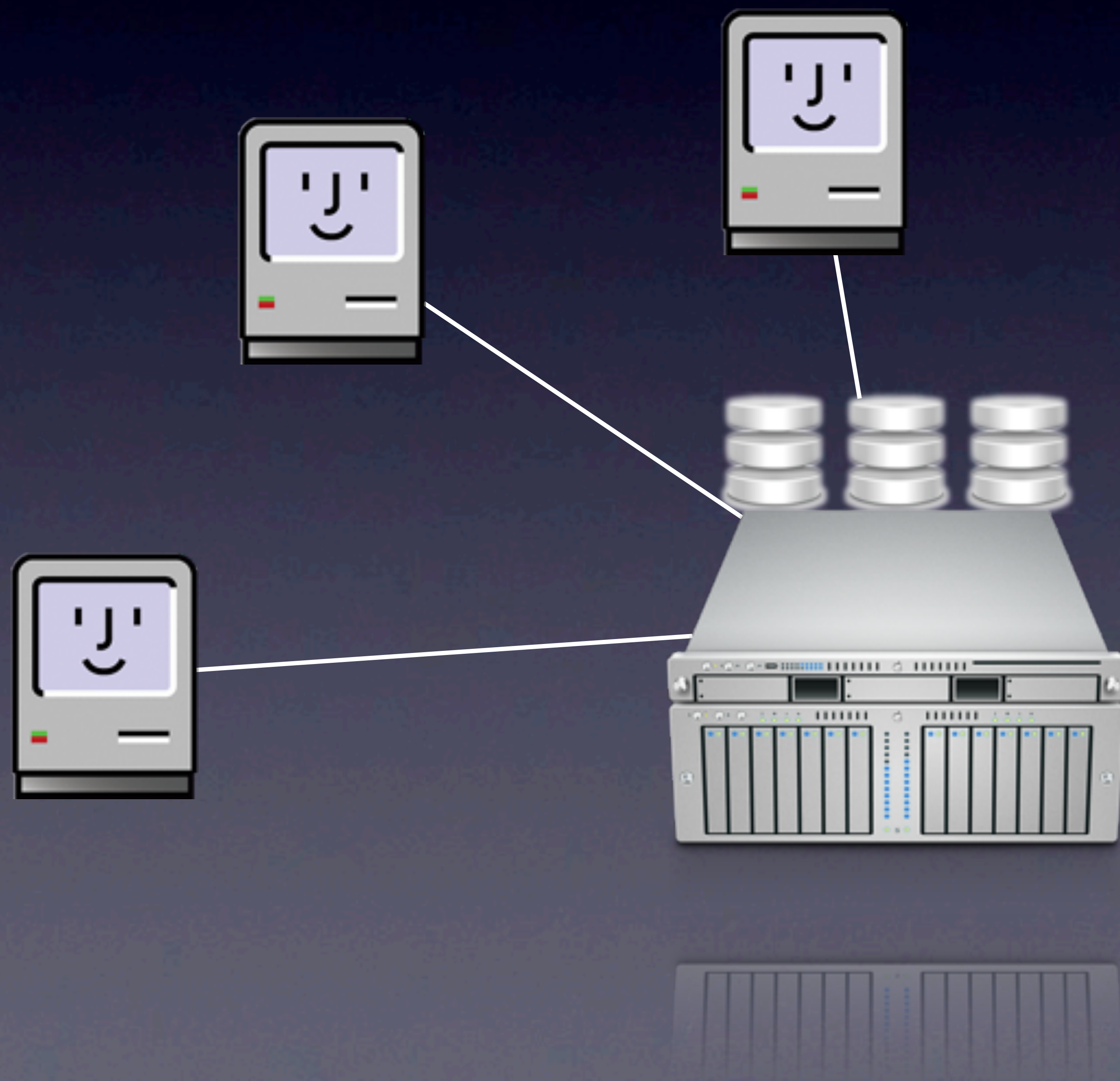
State of the Art Meeting

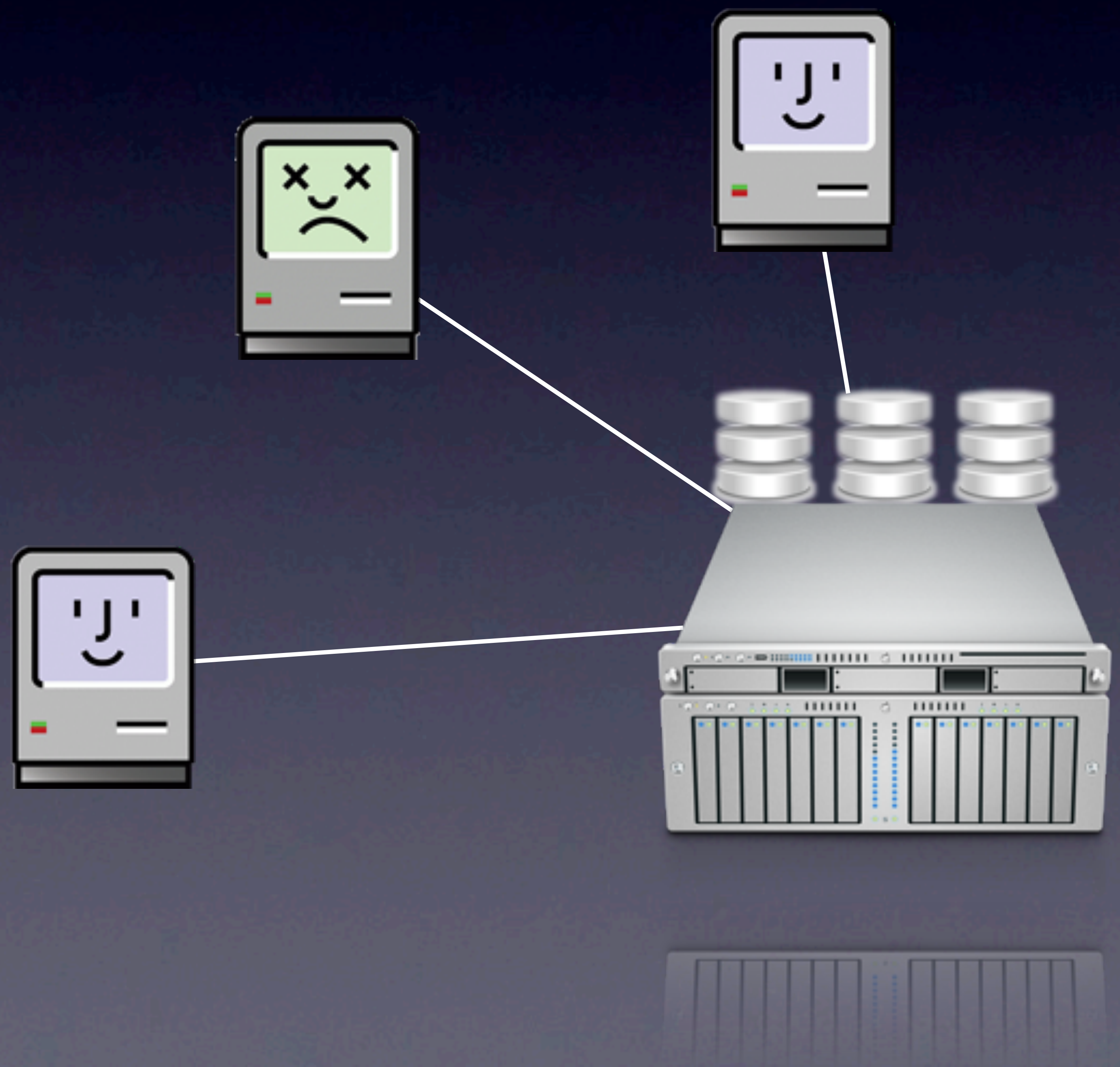
Multi-level Multi-core Distributed Trace Synchronization

Benjamin Poirier

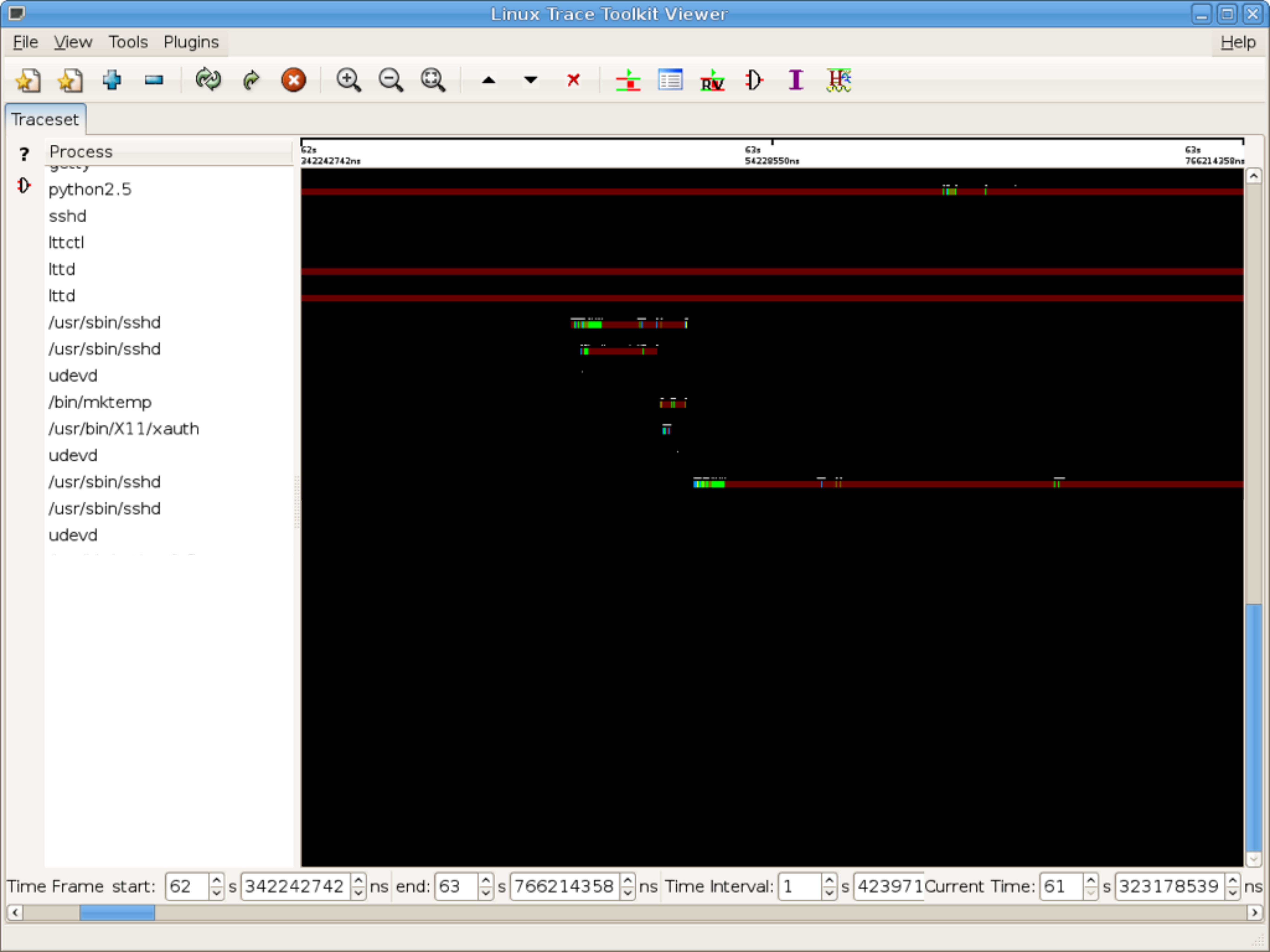
What is Distributed Tracing?

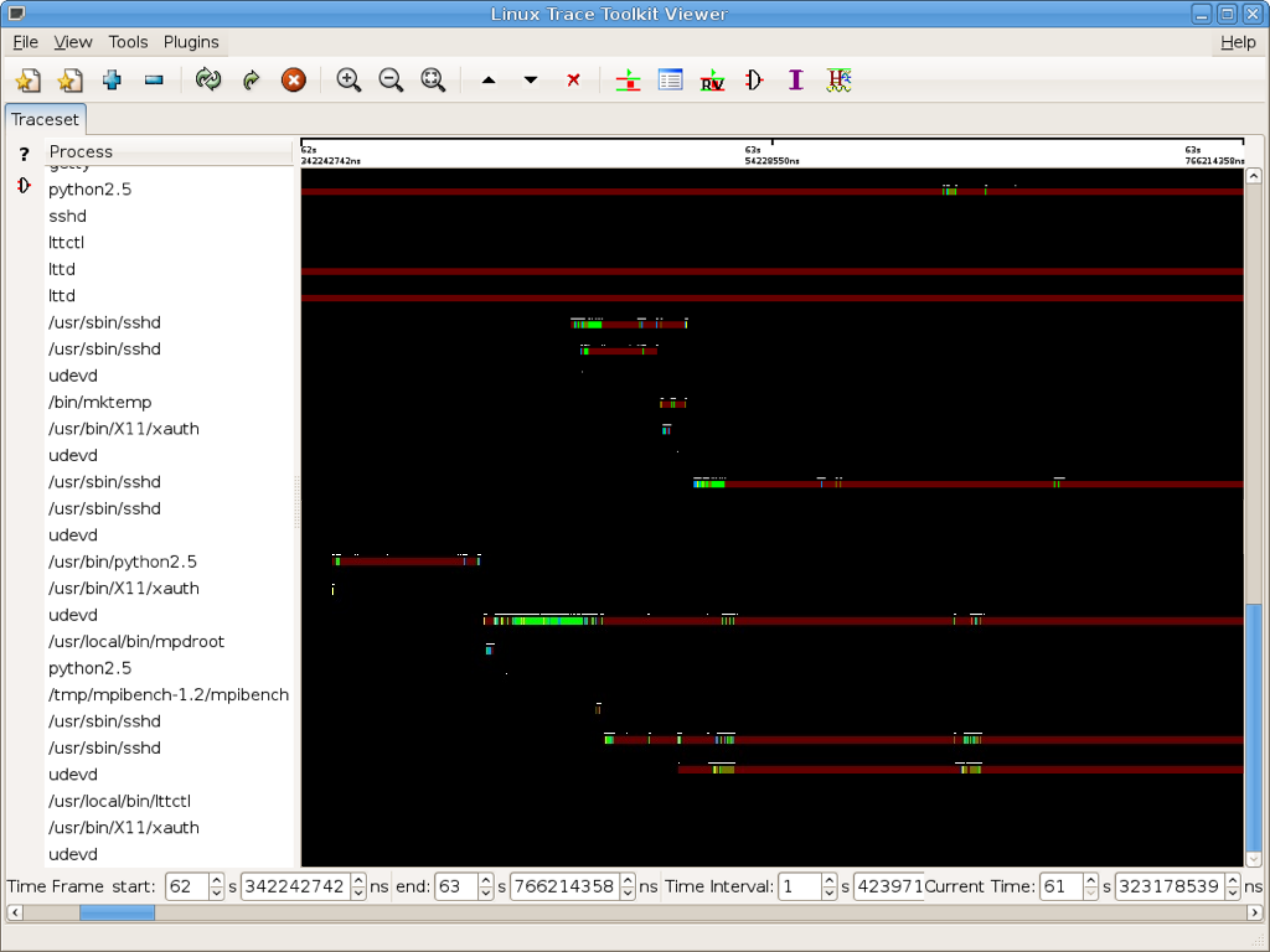


What is Distributed Tracing?

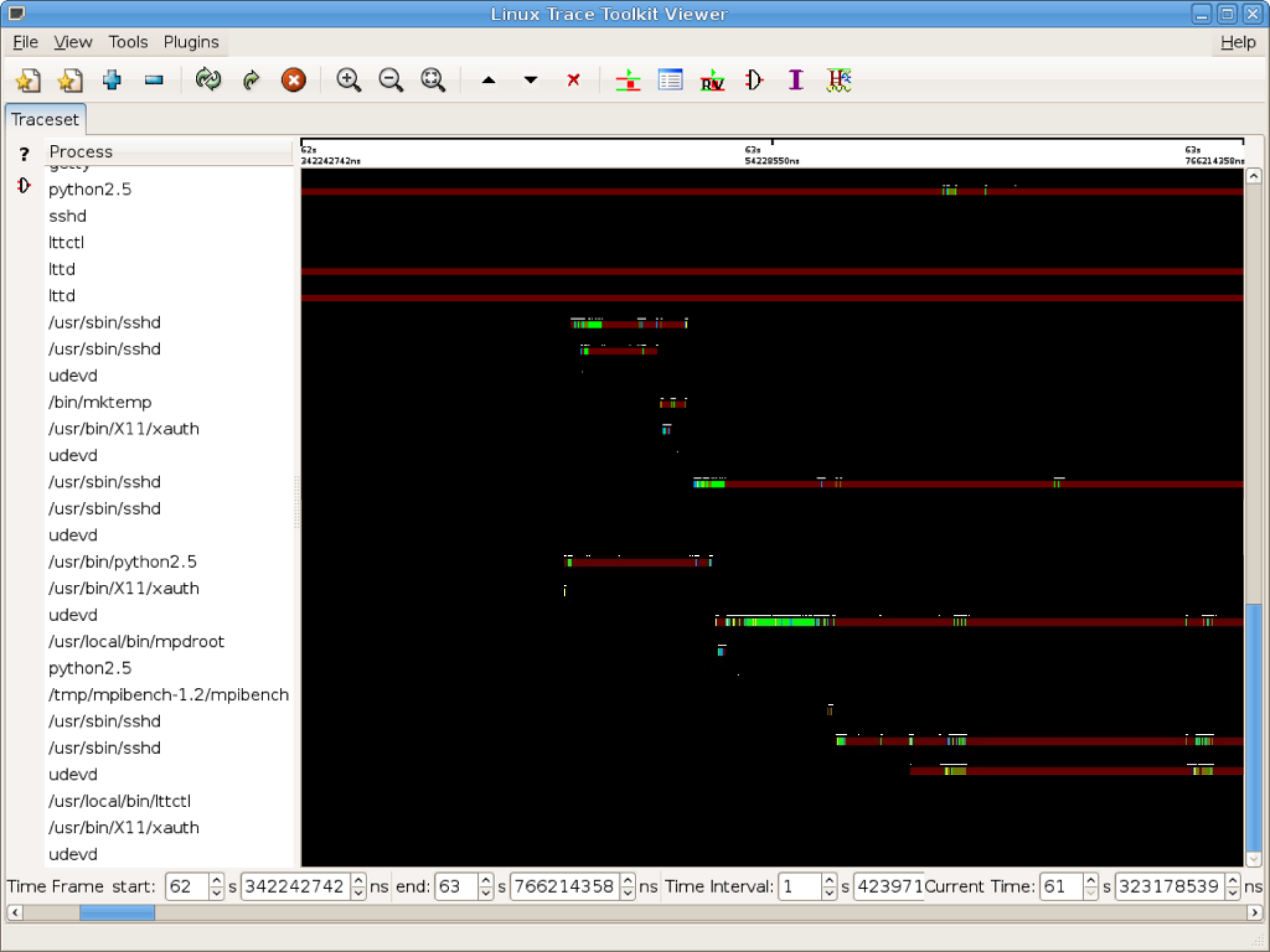


How is distributed tracing done?


Linux Trace Toolkit Viewer
File
View
Tools
Plugins
Help
Traceset
Process
python2.5
sshd
lttctl
lttd
lttd
/usr/sbin/sshd
/usr/sbin/sshd
udevd
/bin/mktemp
/usr/bin/X11/xauth
udevd
/usr/sbin/sshd
/usr/sbin/sshd
udevd
62s
342242742ns
63s
54228550ns
63s
766214358ns
Time Frame start: 62 s 342242742 ns end: 63 s 766214358 ns Time Interval: 1 s 423971
Current Time: 61 s 323178539 ns

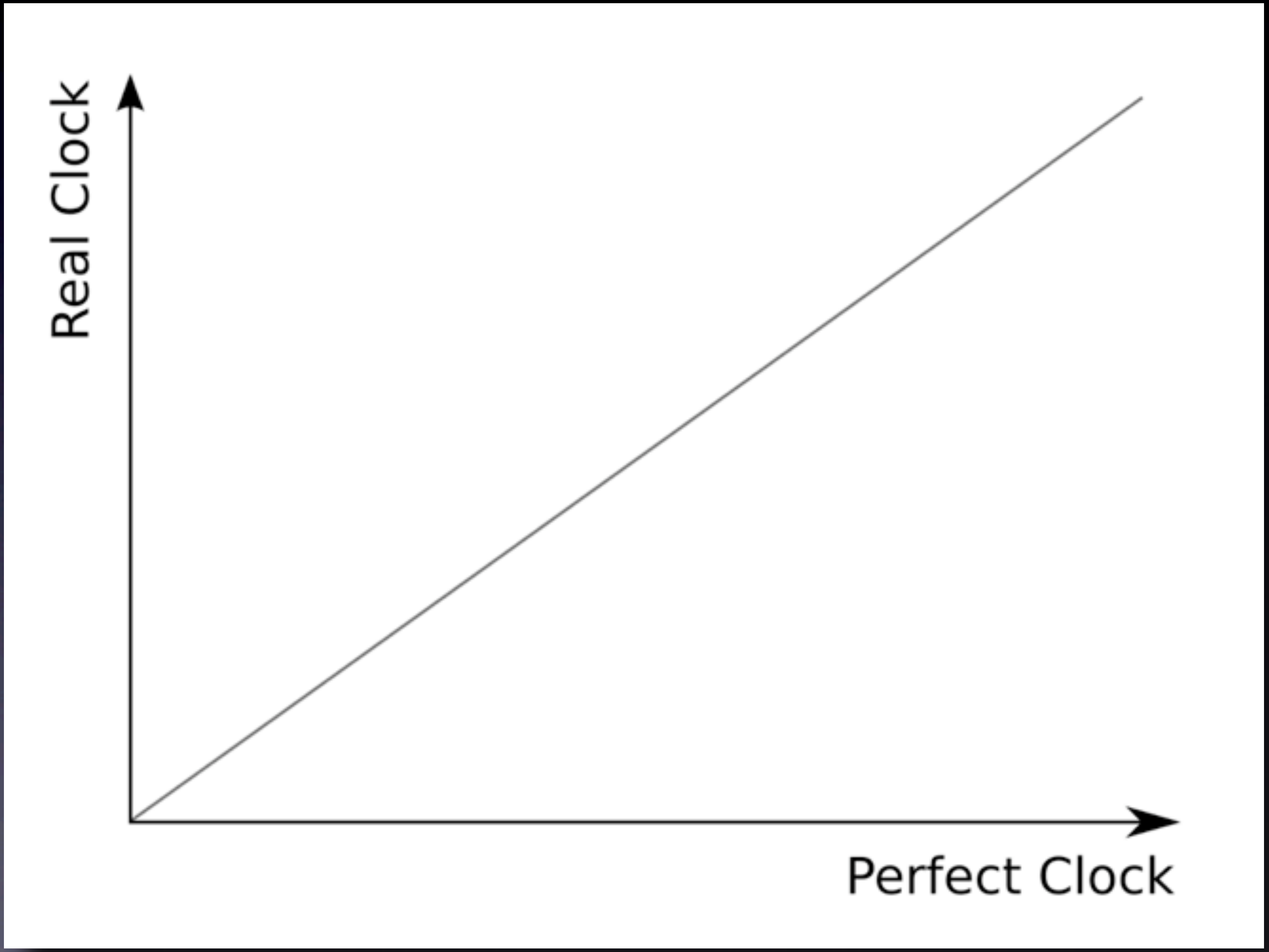


Linux Trace Toolkit Viewer
File View Tools Plugins
Help
Traceset
Process
python2.5
sshd
lttctl
lttd
lttd
/usr/sbin/sshd
/usr/sbin/sshd
udevd
/bin/mktemp
/usr/bin/X11/xauth
udevd
/usr/sbin/sshd
/usr/sbin/sshd
udevd
/usr/bin/python2.5
/usr/bin/X11/xauth
udevd
/usr/local/bin/mpdroot
python2.5
/tmp/mpibench-1.2/mpibench
/usr/sbin/sshd
/usr/sbin/sshd
udevd
/usr/local/bin/lttctl
/usr/bin/X11/xauth
udevd
62s 342242742ns
63s 54228550ns
63s 766214358ns
Time Frame start: 62 s 342242742 ns end: 63 s 766214358 ns Time Interval: 1 s 423971 Current Time: 61 s 323178539 ns


Linux Trace Toolkit Viewer
File
View
Tools
Plugins
Help
Traceset
Process
python2.5
sshd
lttctl
lttd
lttd
/usr/sbin/sshd
/usr/sbin/sshd
udevd
/bin/mktemp
/usr/bin/X11/xauth
udevd
/usr/sbin/sshd
/usr/sbin/sshd
udevd
/usr/bin/python2.5
/usr/bin/X11/xauth
udevd
/usr/local/bin/mpdroot
python2.5
/tmp/mpibench-1.2/mpibench
/usr/sbin/sshd
/usr/sbin/sshd
udevd
/usr/local/bin/lttctl
/usr/bin/X11/xauth
udevd
Time Frame start: 62 s 342242742 ns end: 63 s 766214358 ns Time Interval: 1 s 423971 Current Time: 61 s 323178539 ns

How is distributed tracing done?

Clock parameters

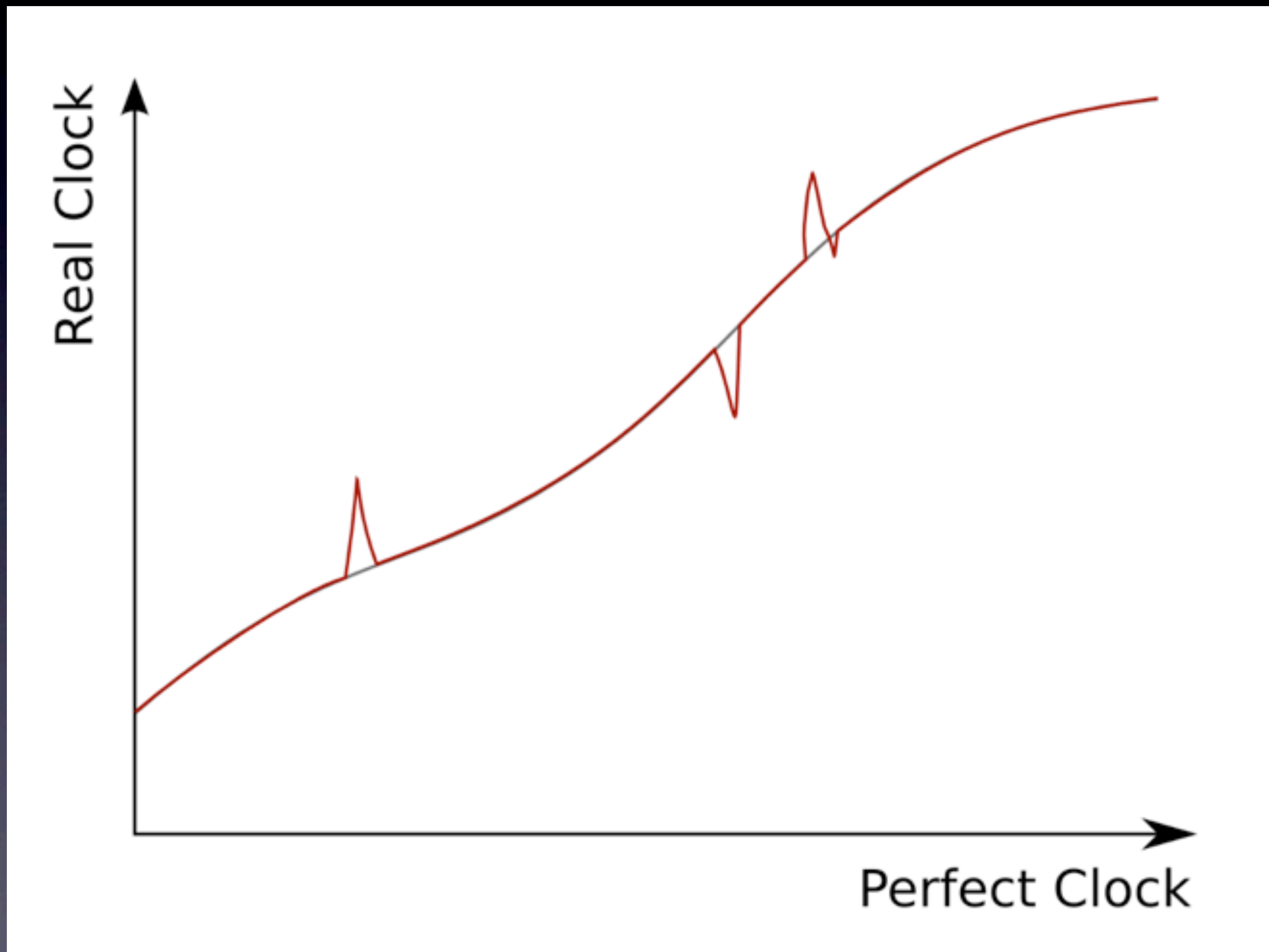


Clock parameters

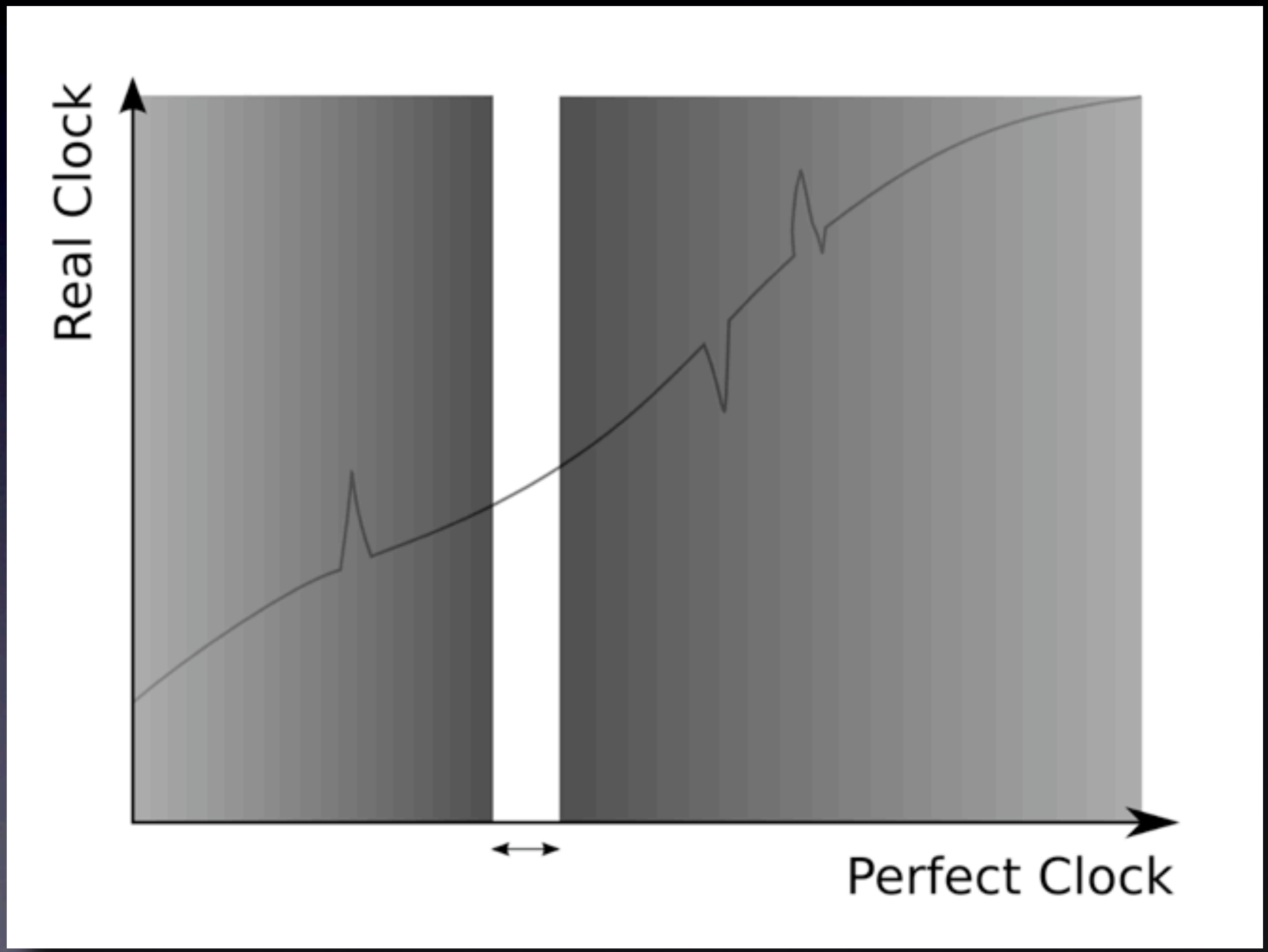
Clock parameters

Clock parameters

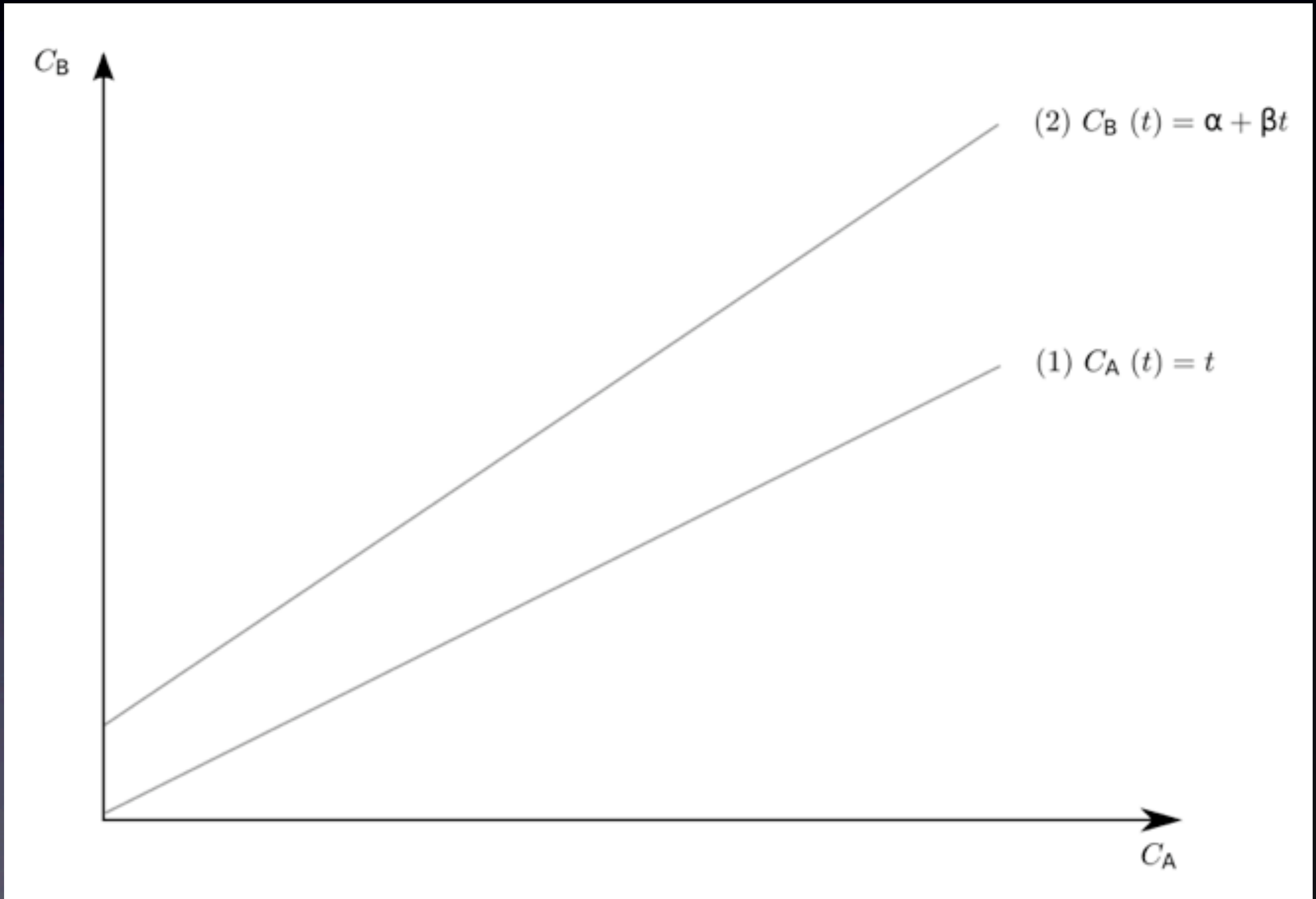
Clock parameters



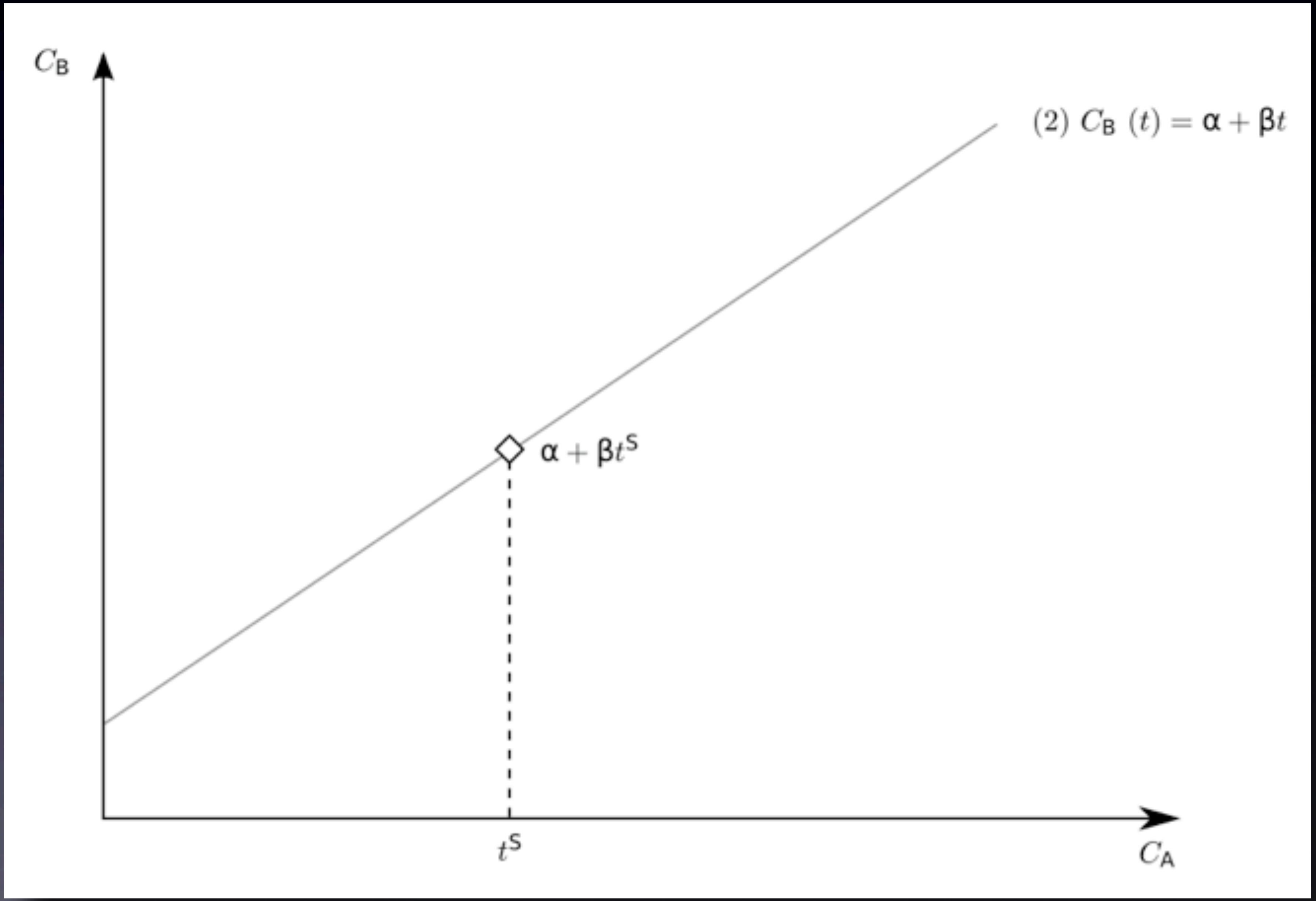
Clock parameters



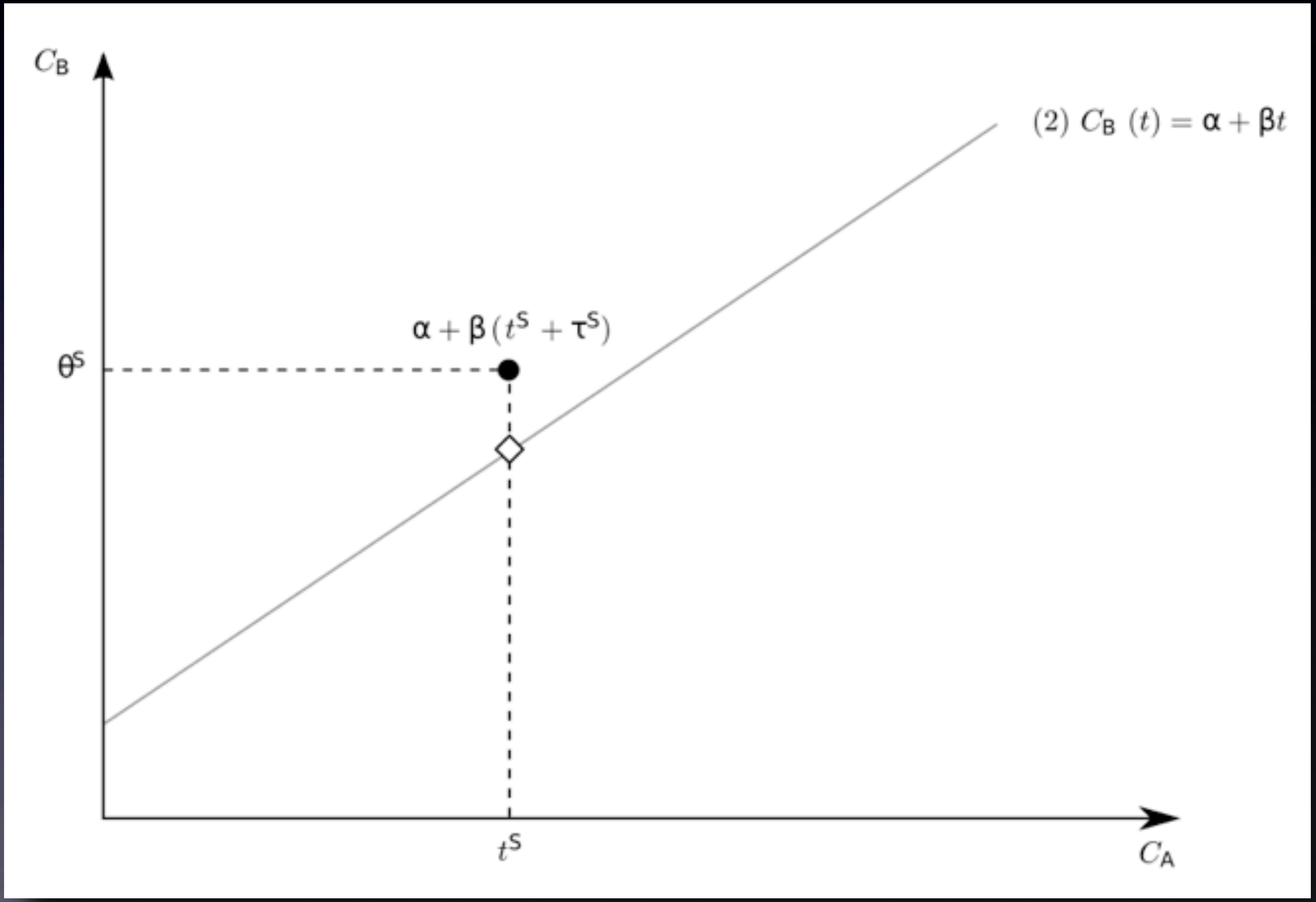
Packet ordering



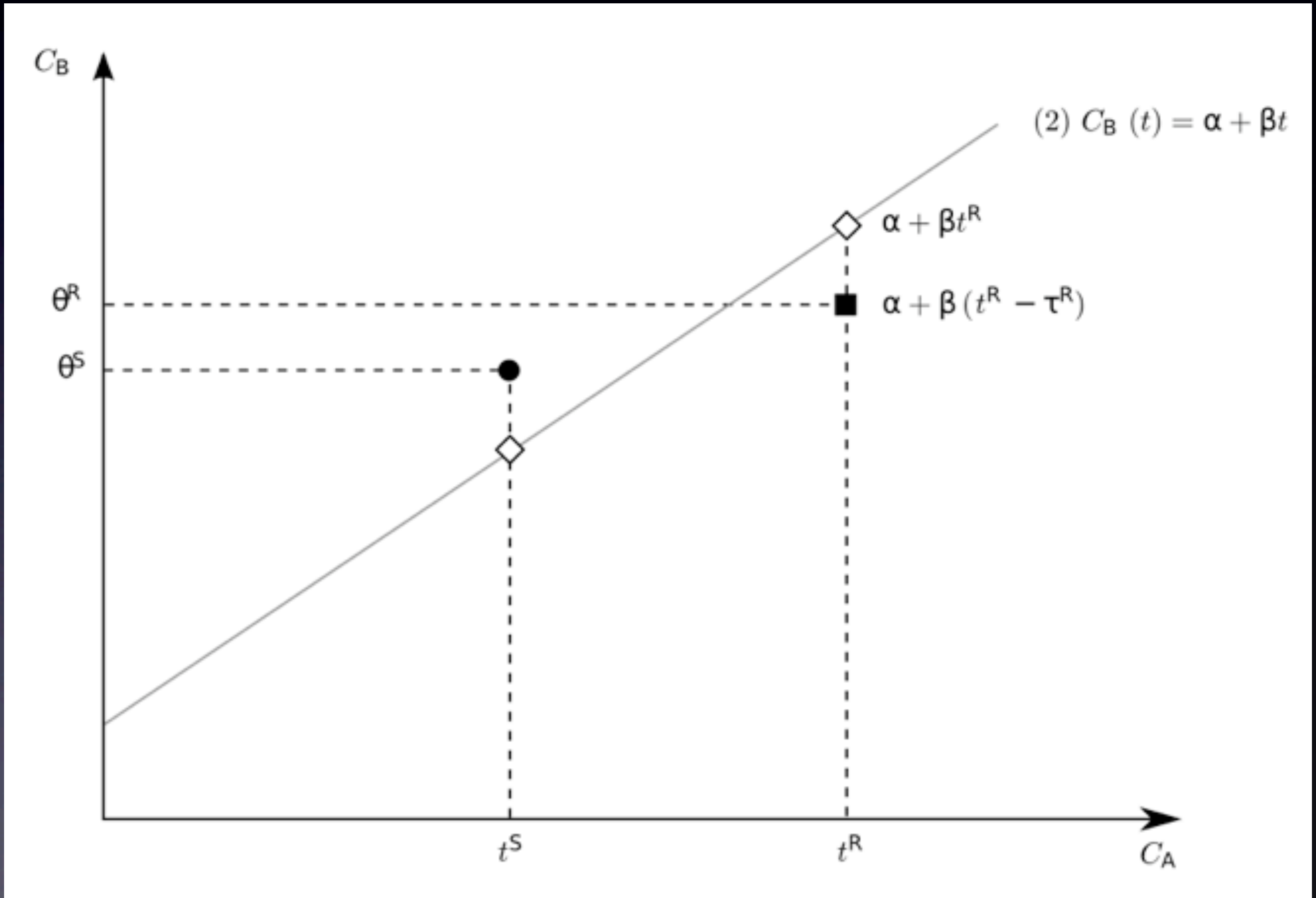
Packet ordering



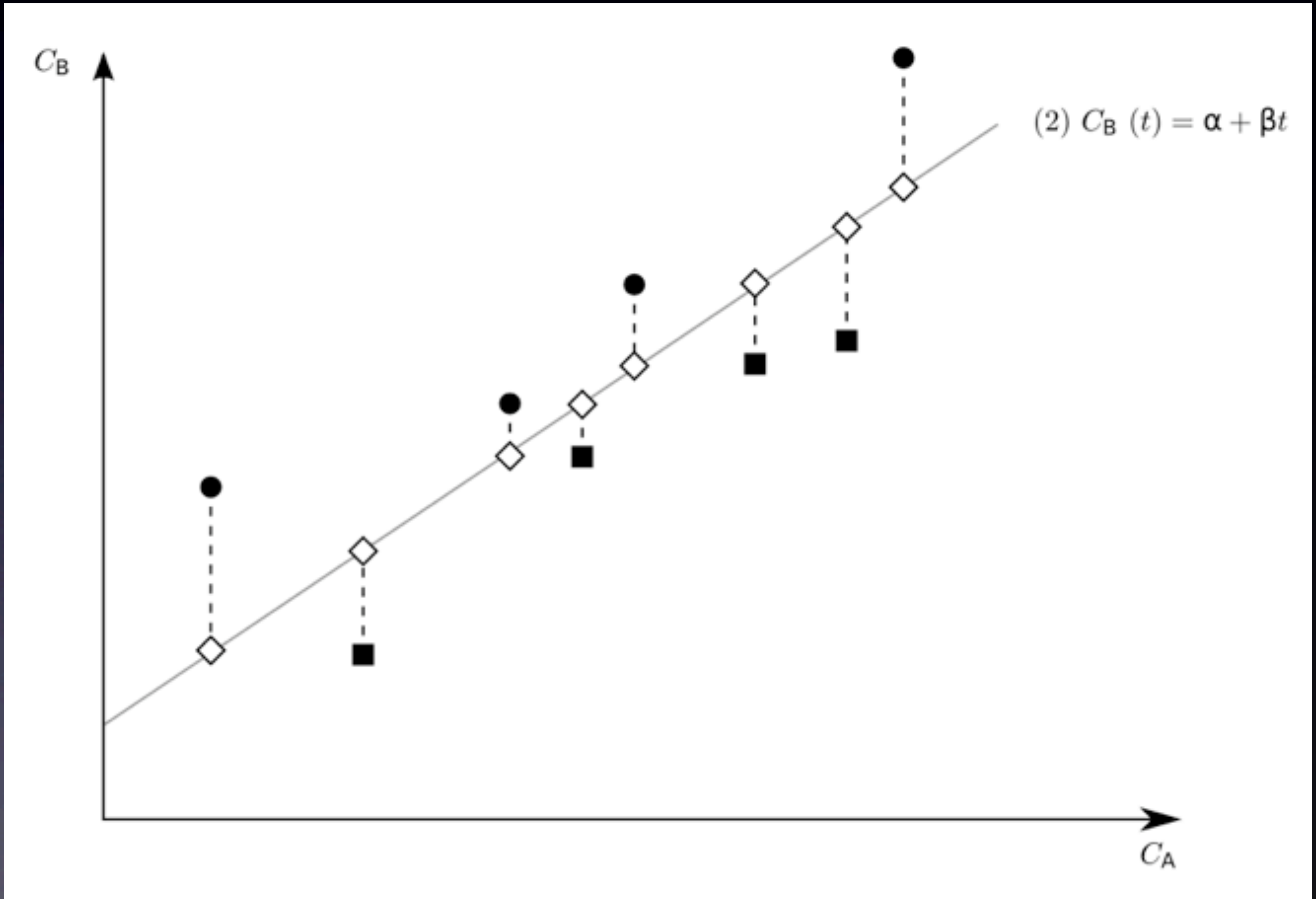
Packet ordering



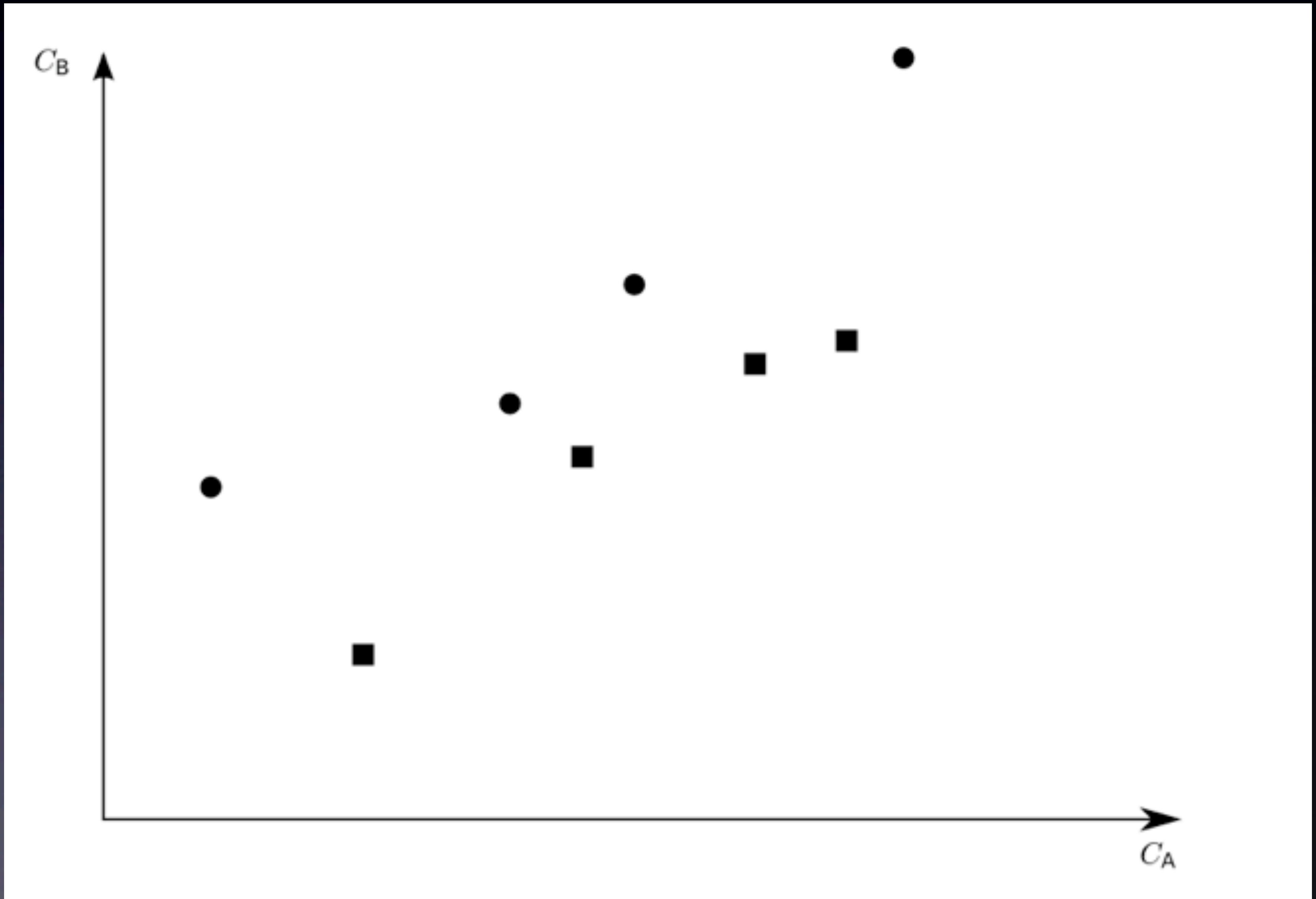
Packet ordering



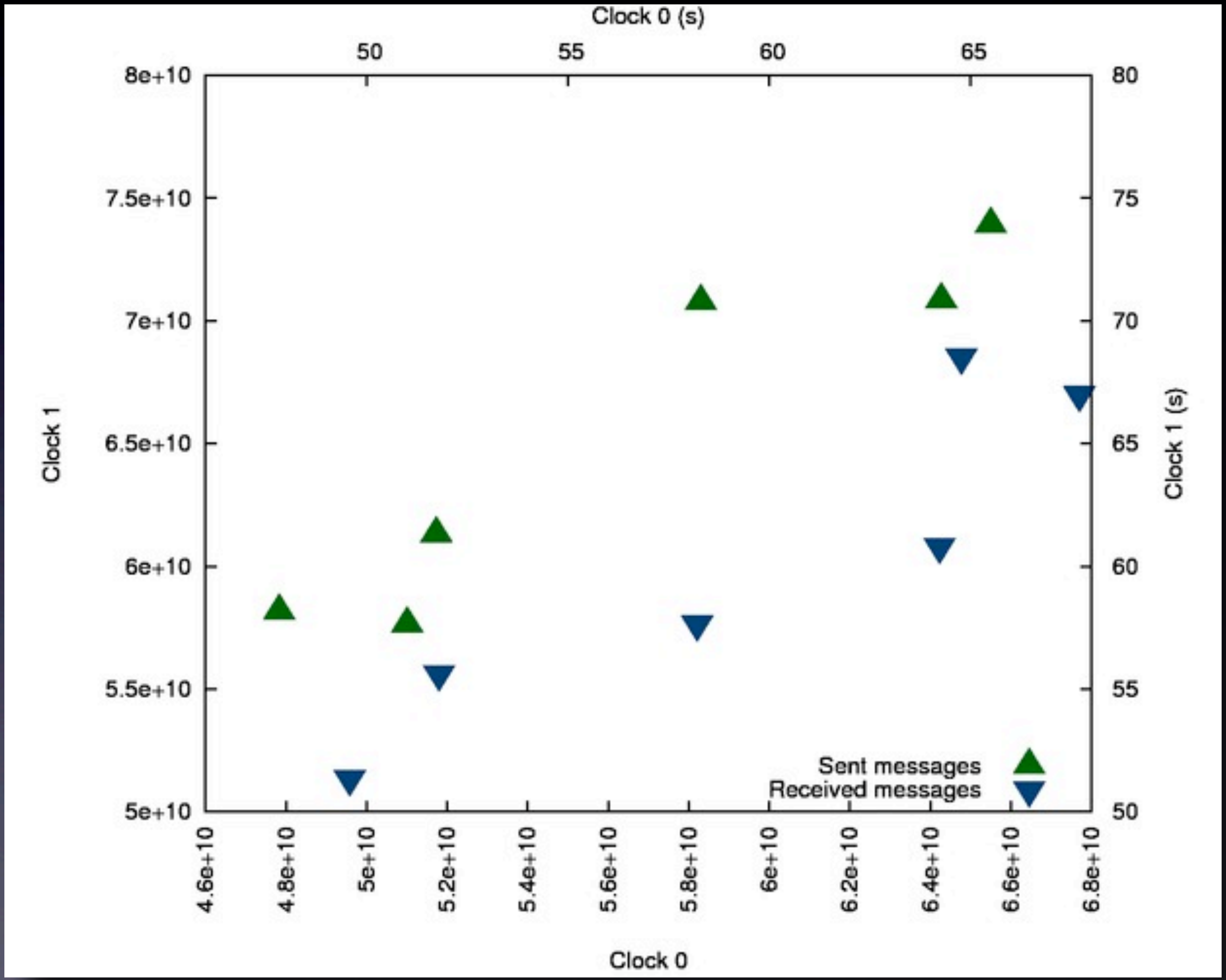
Packet ordering

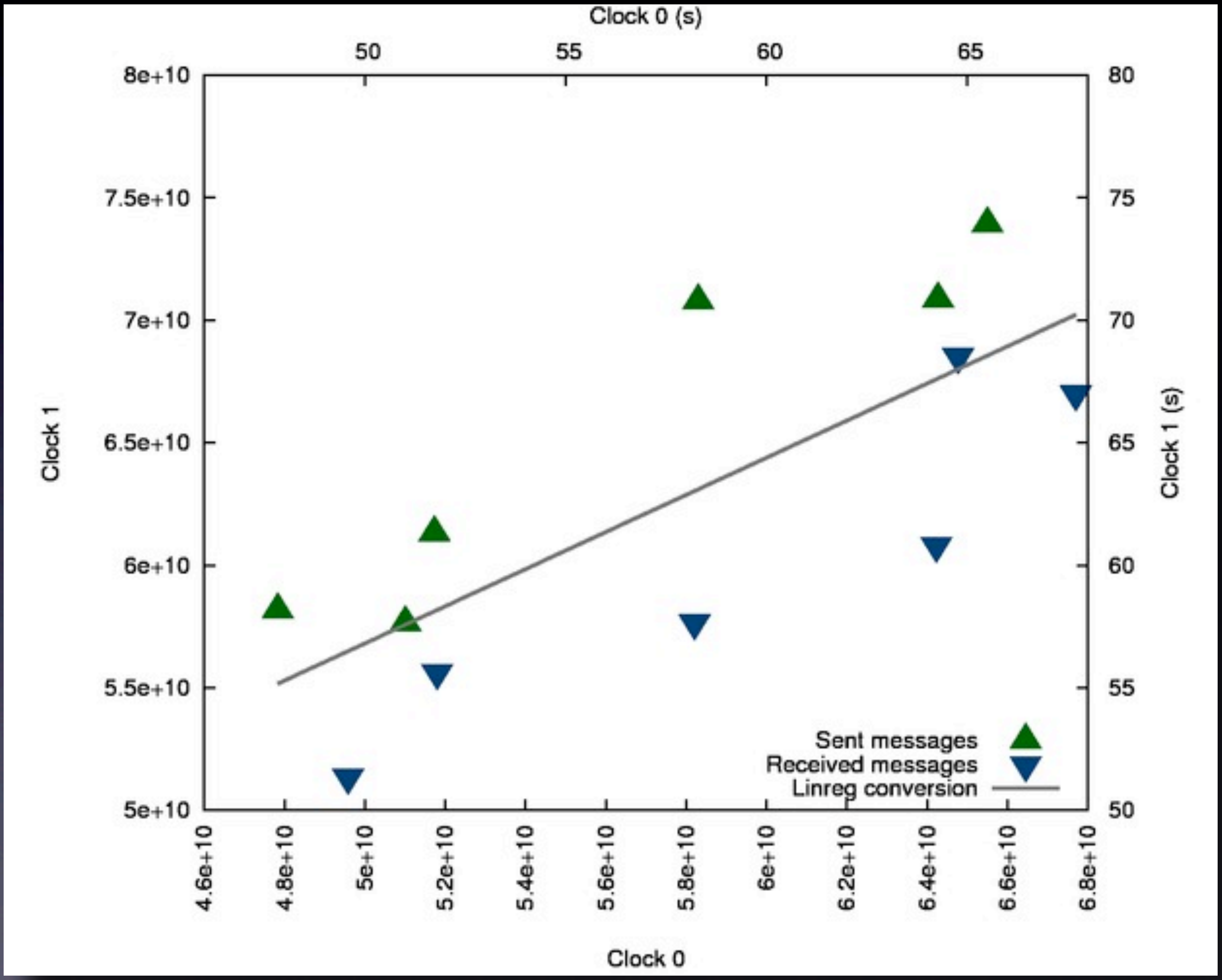


Packet ordering

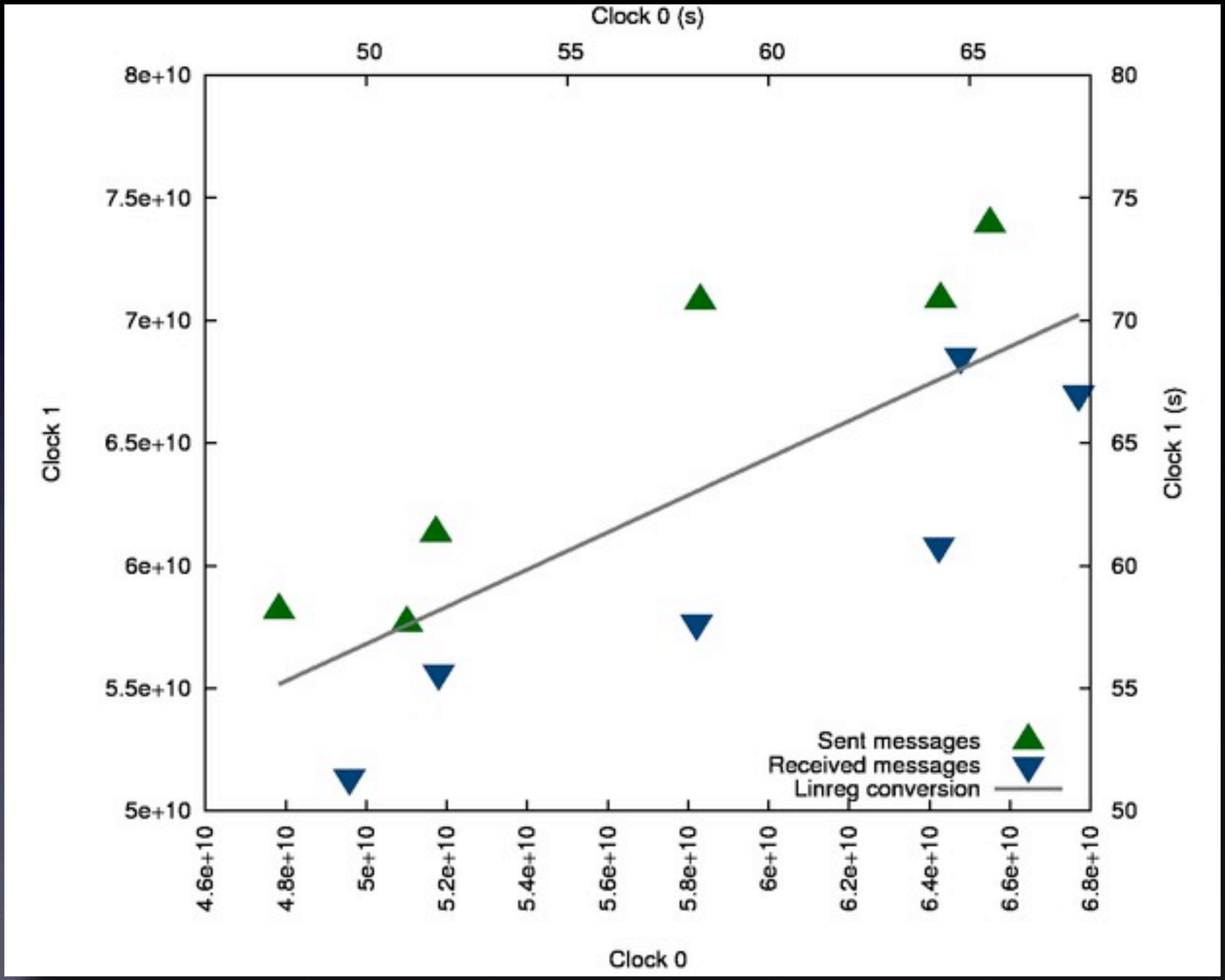


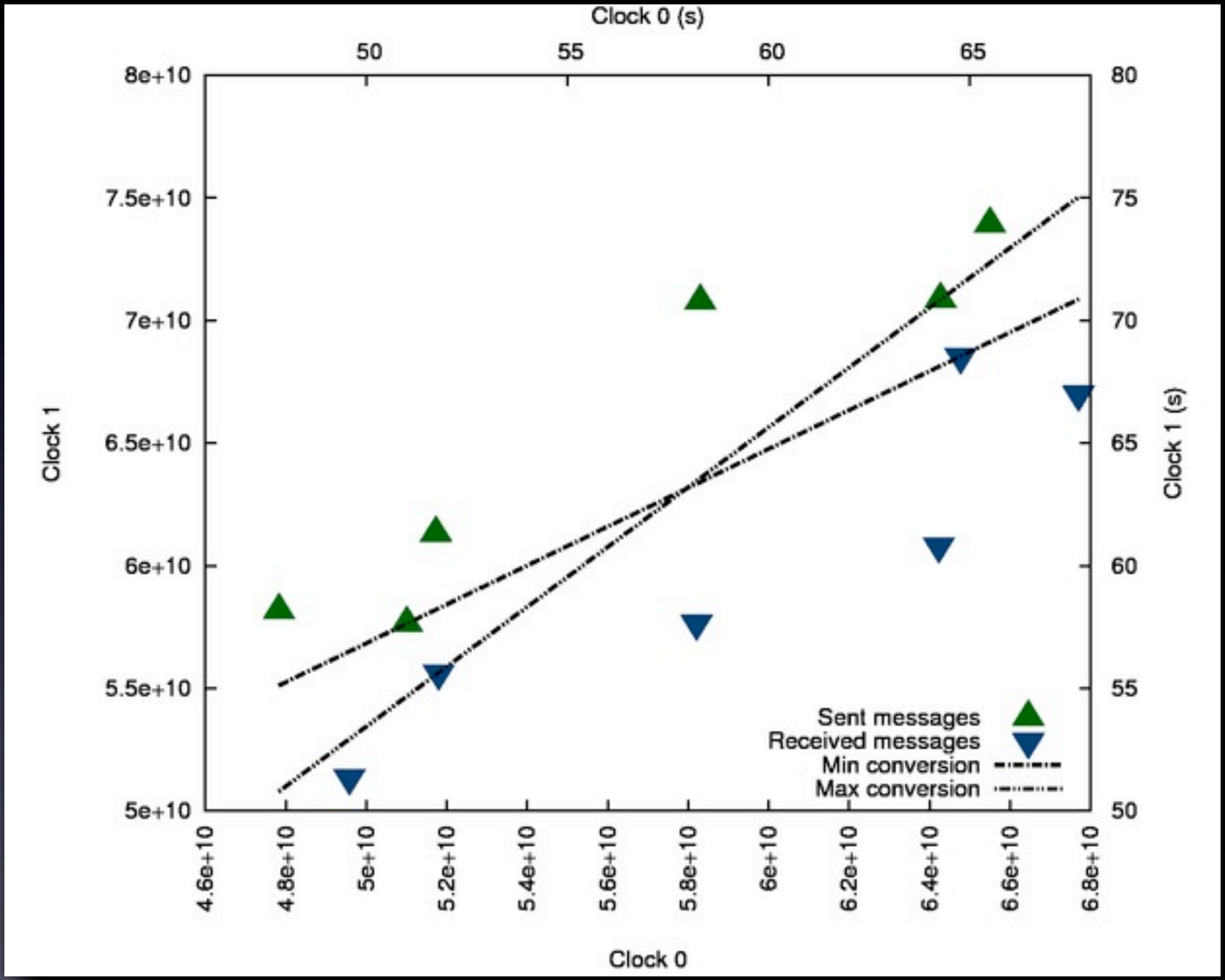
Correction factors



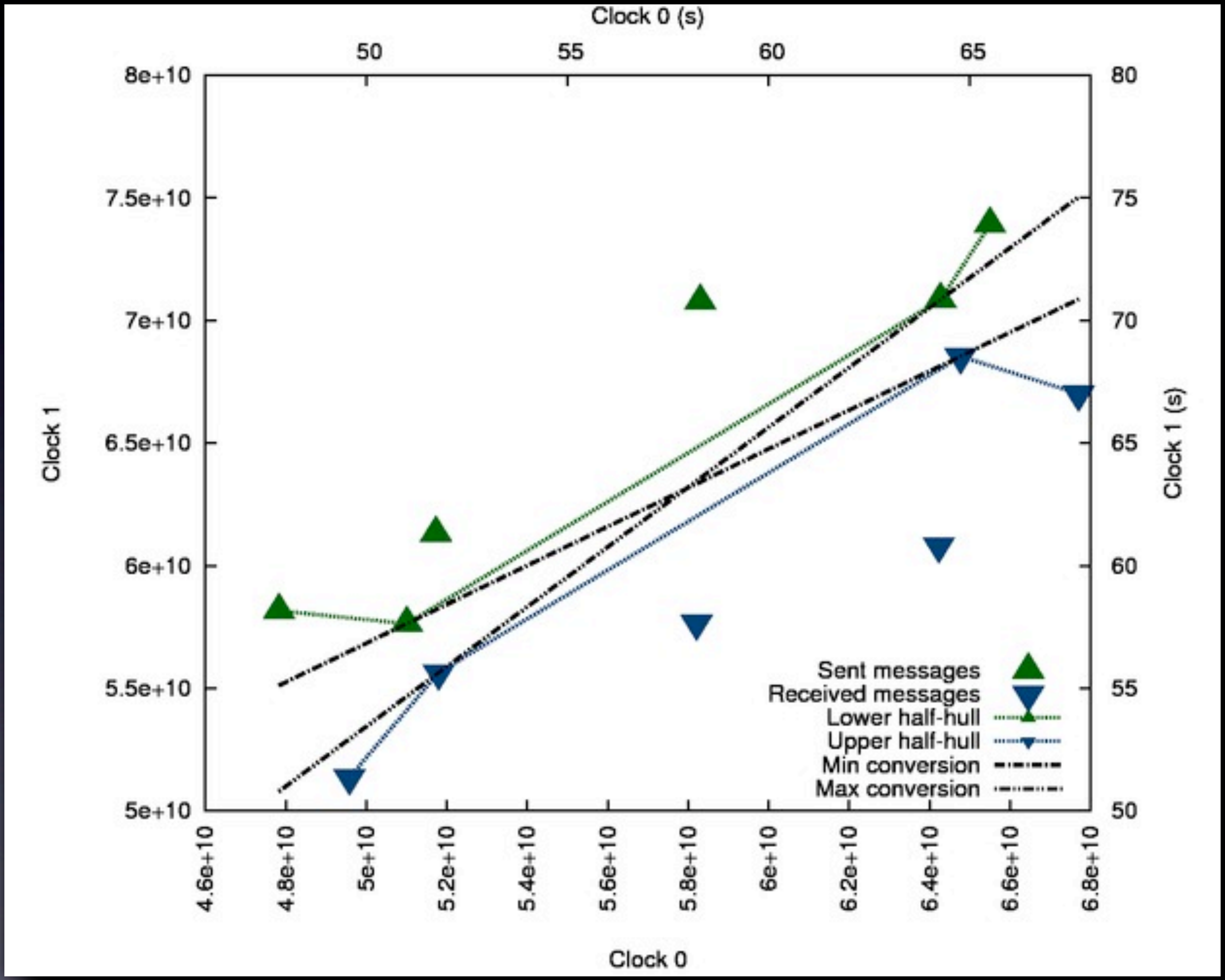

Clock 0 (s)
50
55
60
65
8e+10
7.5e+10
7e+10
6.5e+10
6e+10
5.5e+10
5e+10
Clock 1
80
75
70
65
60
55
50
Clock 1 (s)
Sent messages
Received messages
Linreg conversion
4.6e+10
4.8e+10
5e+10
5.2e+10
5.4e+10
5.6e+10
5.8e+10
6e+10
6.2e+10
6.4e+10
6.6e+10
6.8e+10
Clock 0

Linear Regression Method



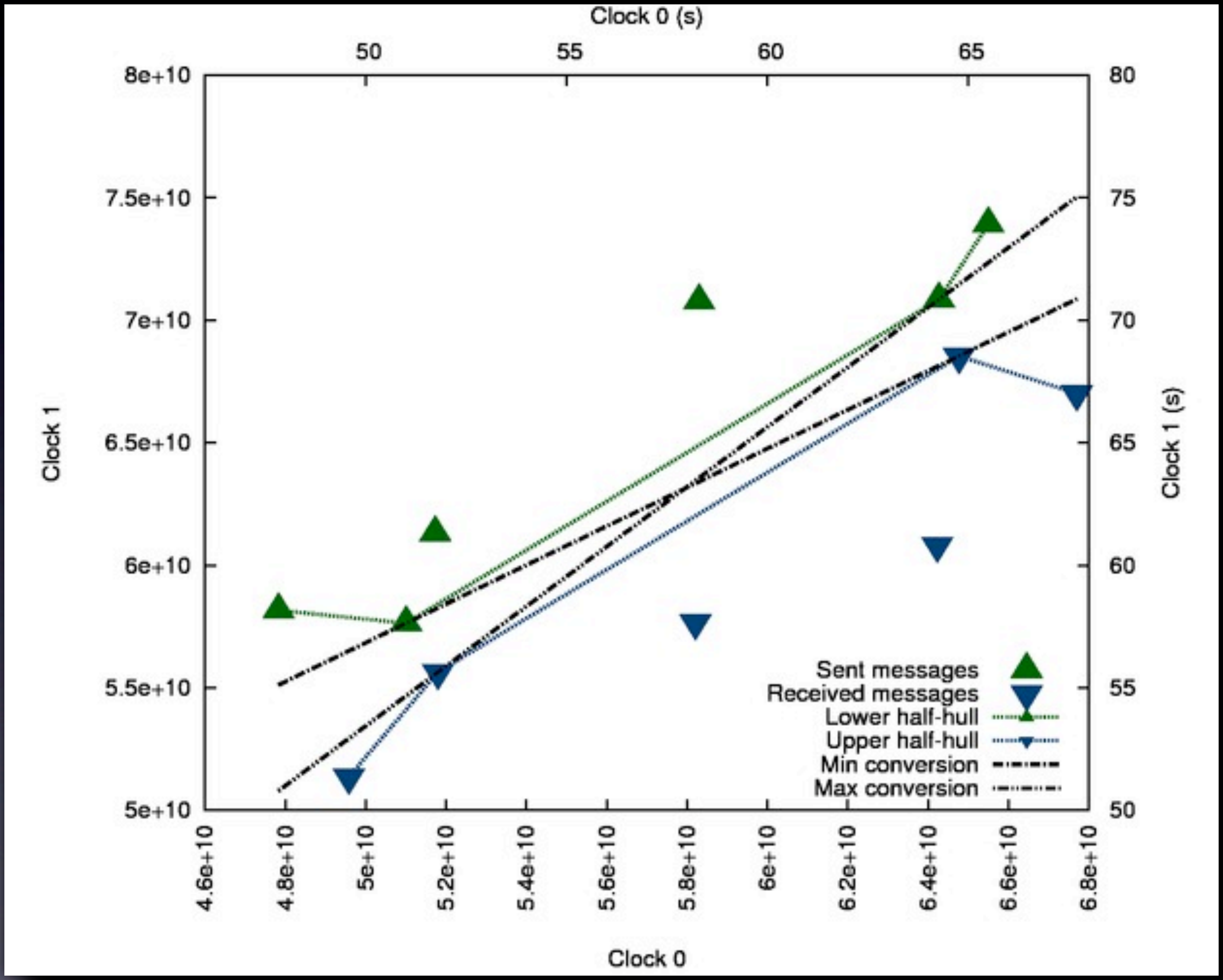


Clock 0 (s)
50
55
60
65
8e+10
7.5e+10
7e+10
6.5e+10
6e+10
5.5e+10
5e+10
Clock 1
80
75
70
65
60
55
50
Clock 1 (s)
Sent messages
Received messages
Min conversion
Max conversion
4.6e+10
4.8e+10
5e+10
5.2e+10
5.4e+10
5.6e+10
5.8e+10
6e+10
6.2e+10
6.4e+10
6.6e+10
6.8e+10
Clock 0

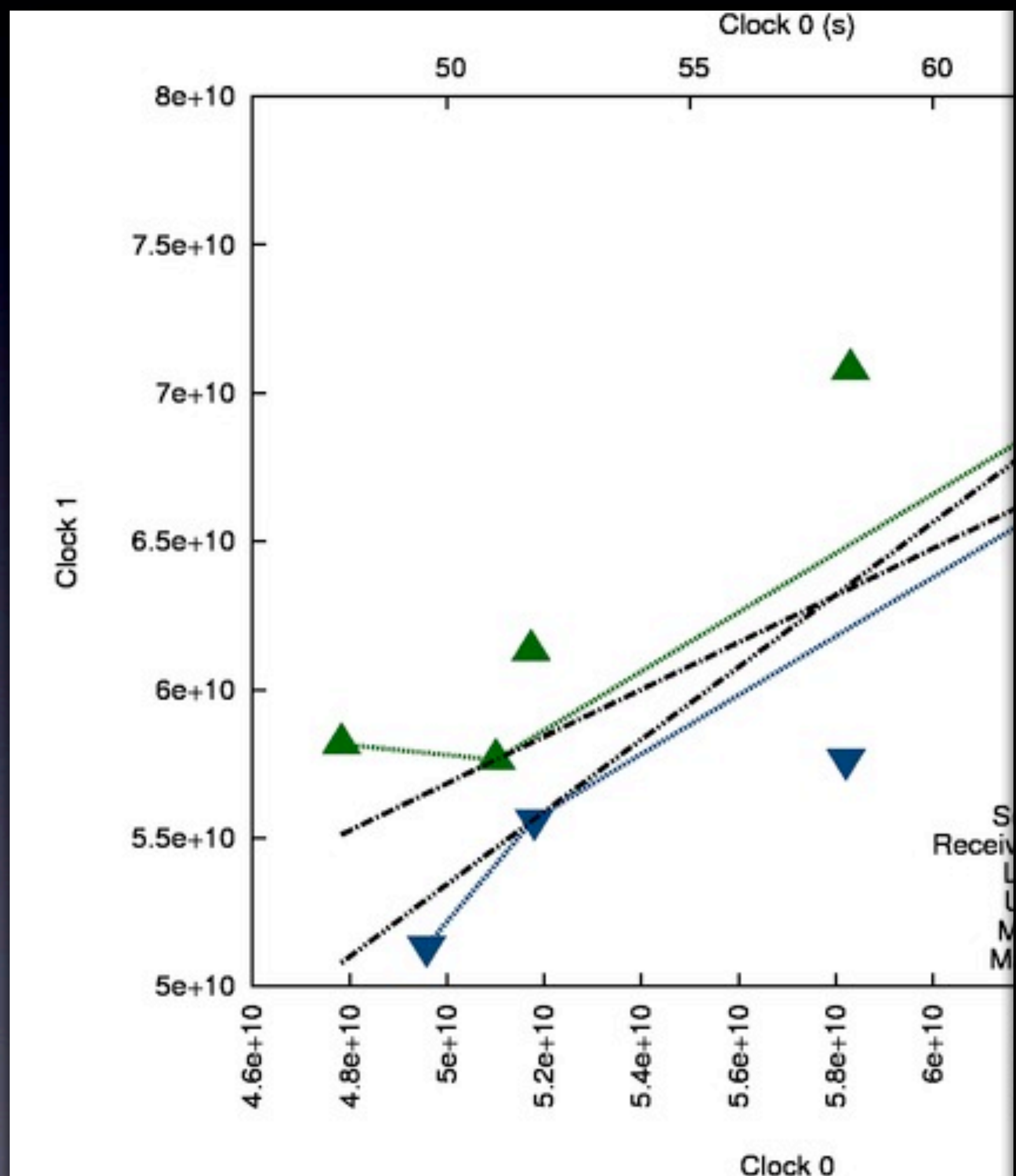


Clock 0 (s)
50
55
60
65
8e+10
7.5e+10
7e+10
6.5e+10
6e+10
5.5e+10
5e+10
80
75
70
65
60
55
50
Clock 1
Clock 1 (s)
Sent messages
Received messages
Lower half-hull
Upper half-hull
Min conversion
Max conversion
4.6e+10
4.8e+10
5e+10
5.2e+10
5.4e+10
5.6e+10
5.8e+10
6e+10
6.2e+10
6.4e+10
6.6e+10
6.8e+10
Clock 0

Convex Hull Method



Convex Hull Method



Optimize:

{minimize, maximize} β

Subject to:

$\alpha+\beta\, x1 \leq y1$

$\alpha+\beta\, x2 \leq y2$

...

$\alpha+\beta\, x5 \geq y5$

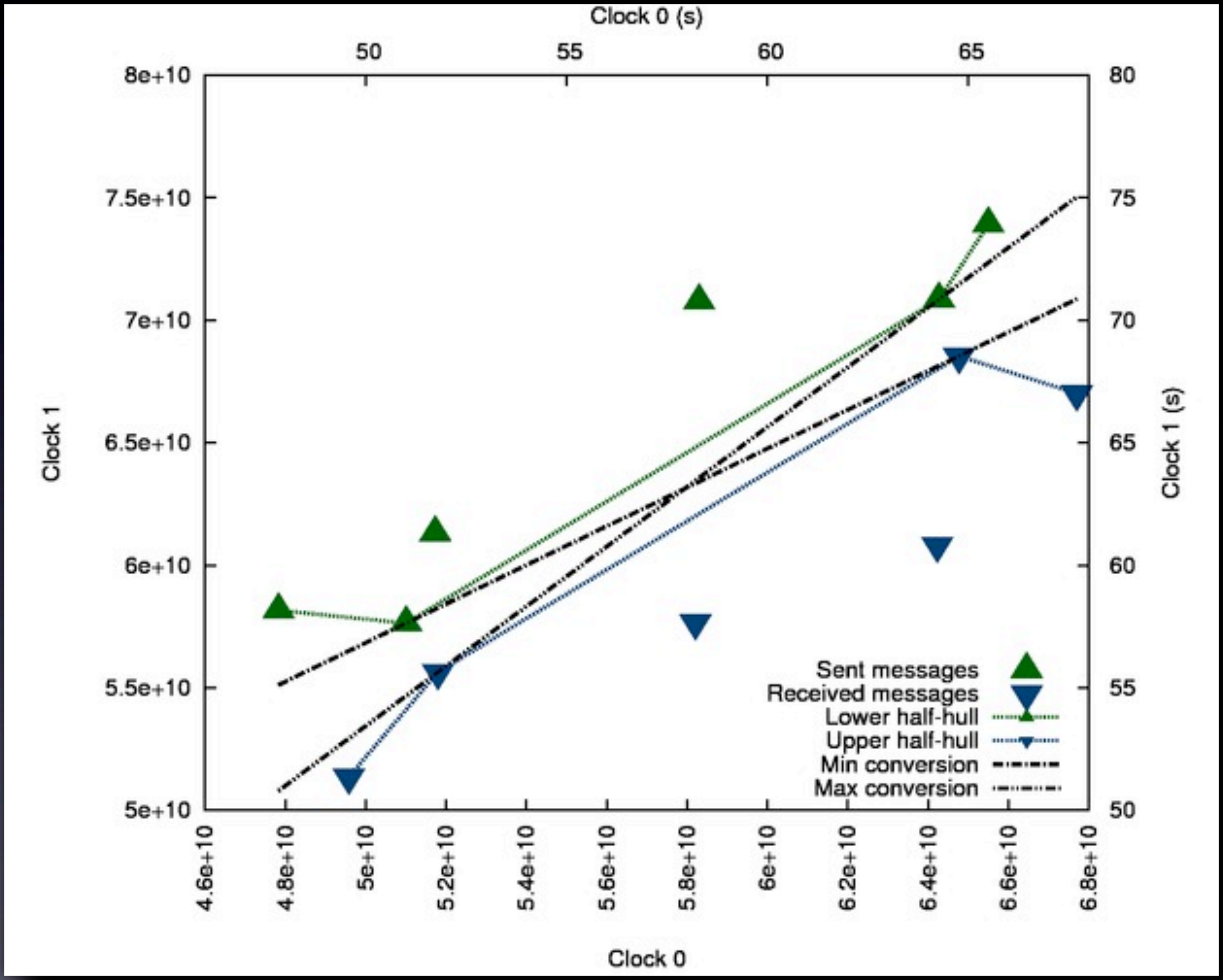
...

With bounds:

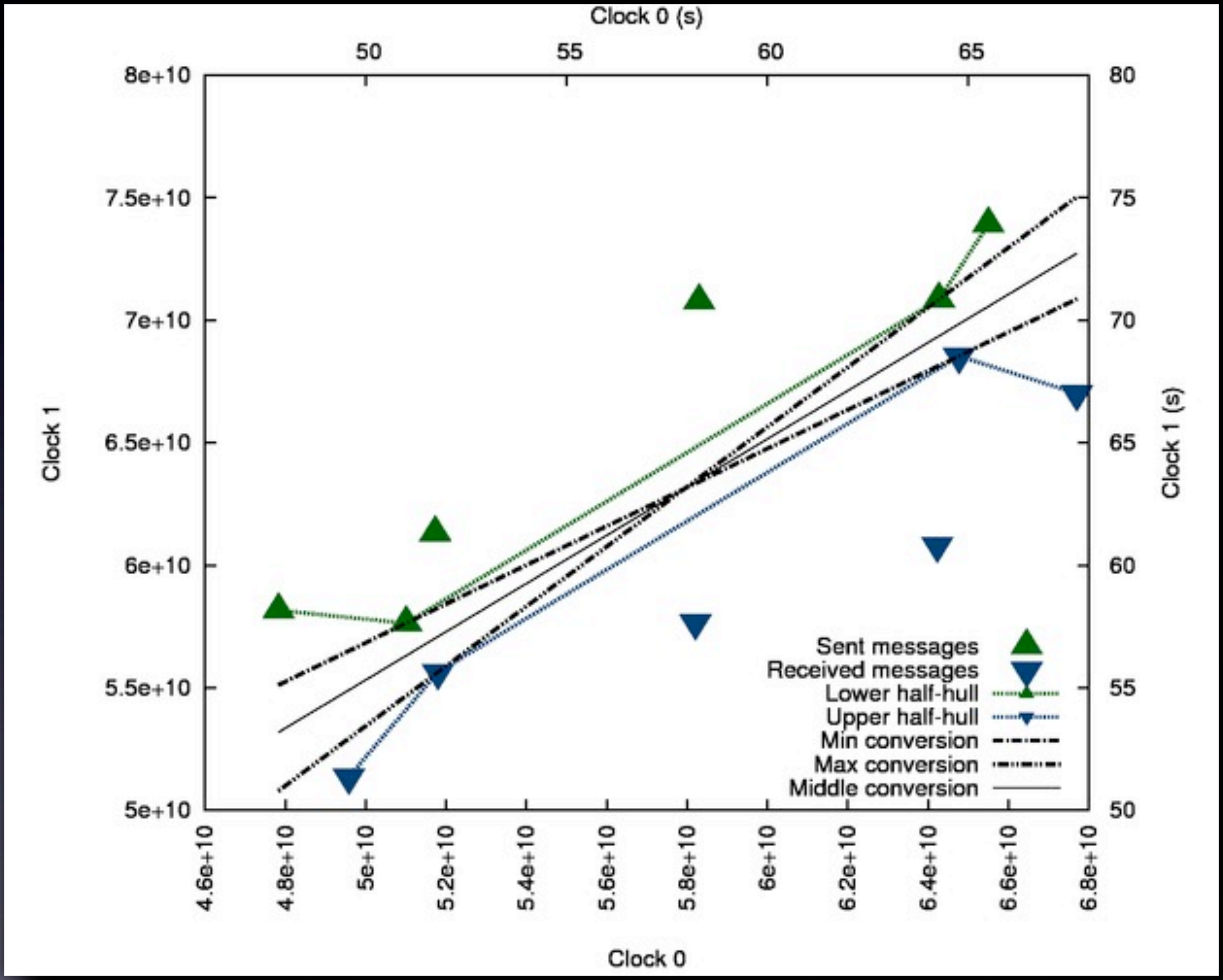
$-\infty \leq \alpha \leq \infty$

$0 \leq \beta \leq \infty$

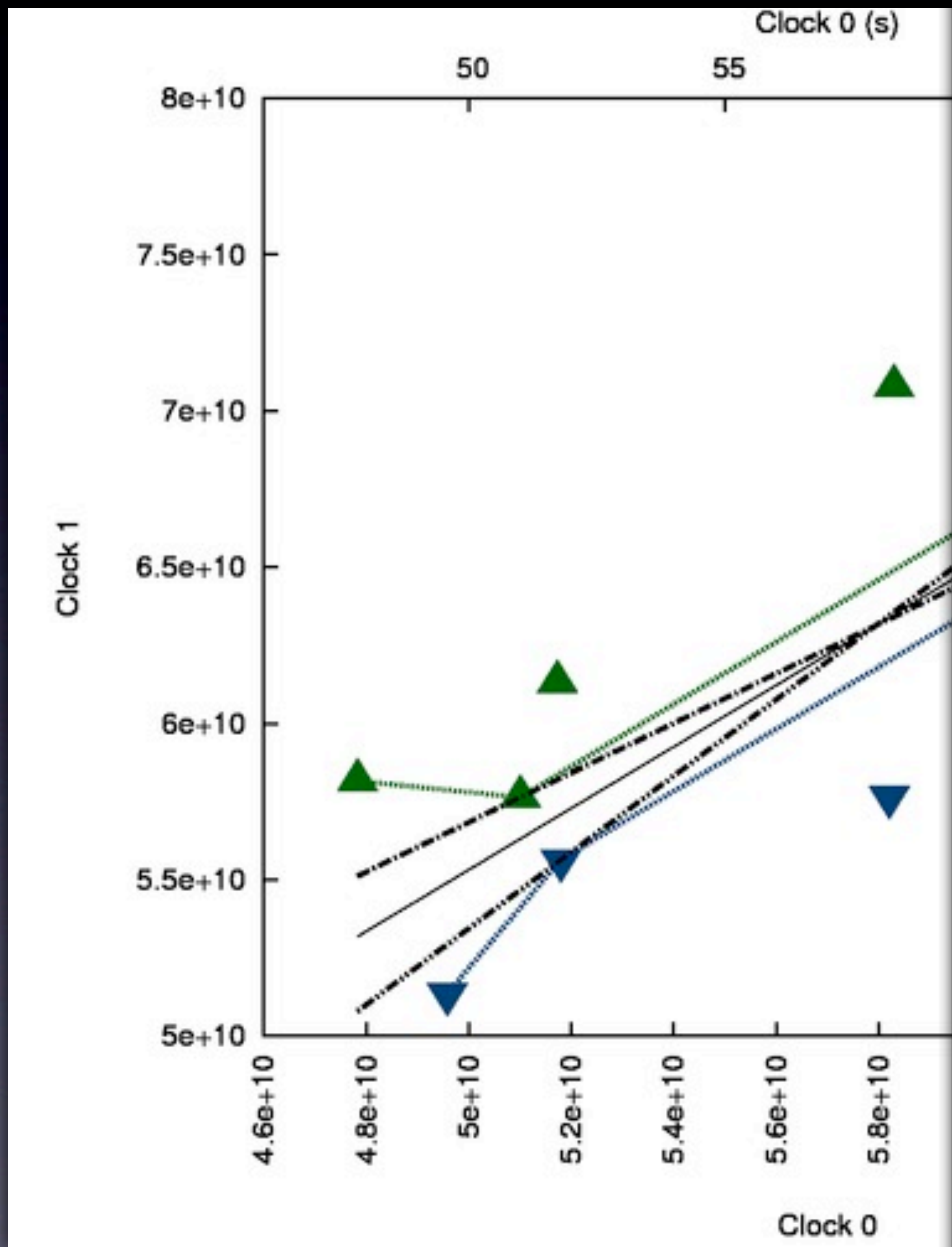
Convex Hull Method



Convex Hull Method



Convex Hull Method



Optimize:

{minimize, maximize} $\alpha+\beta\, t$

Subject to:

$\alpha+\beta\, x1 \leq y1$

$\alpha+\beta\, x2 \leq y2$

...

$\alpha+\beta\, x5 \geq y5$

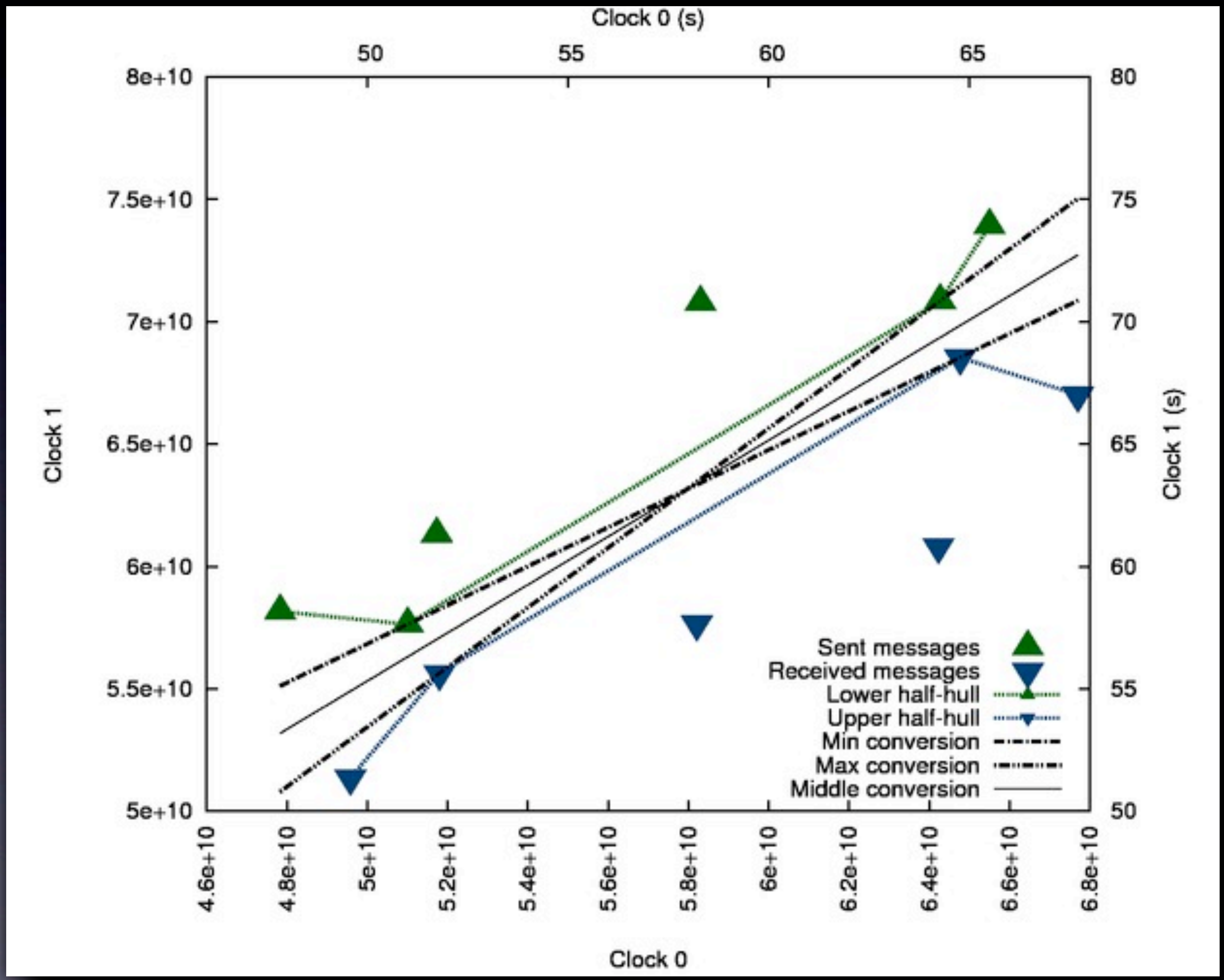
...

With bounds:

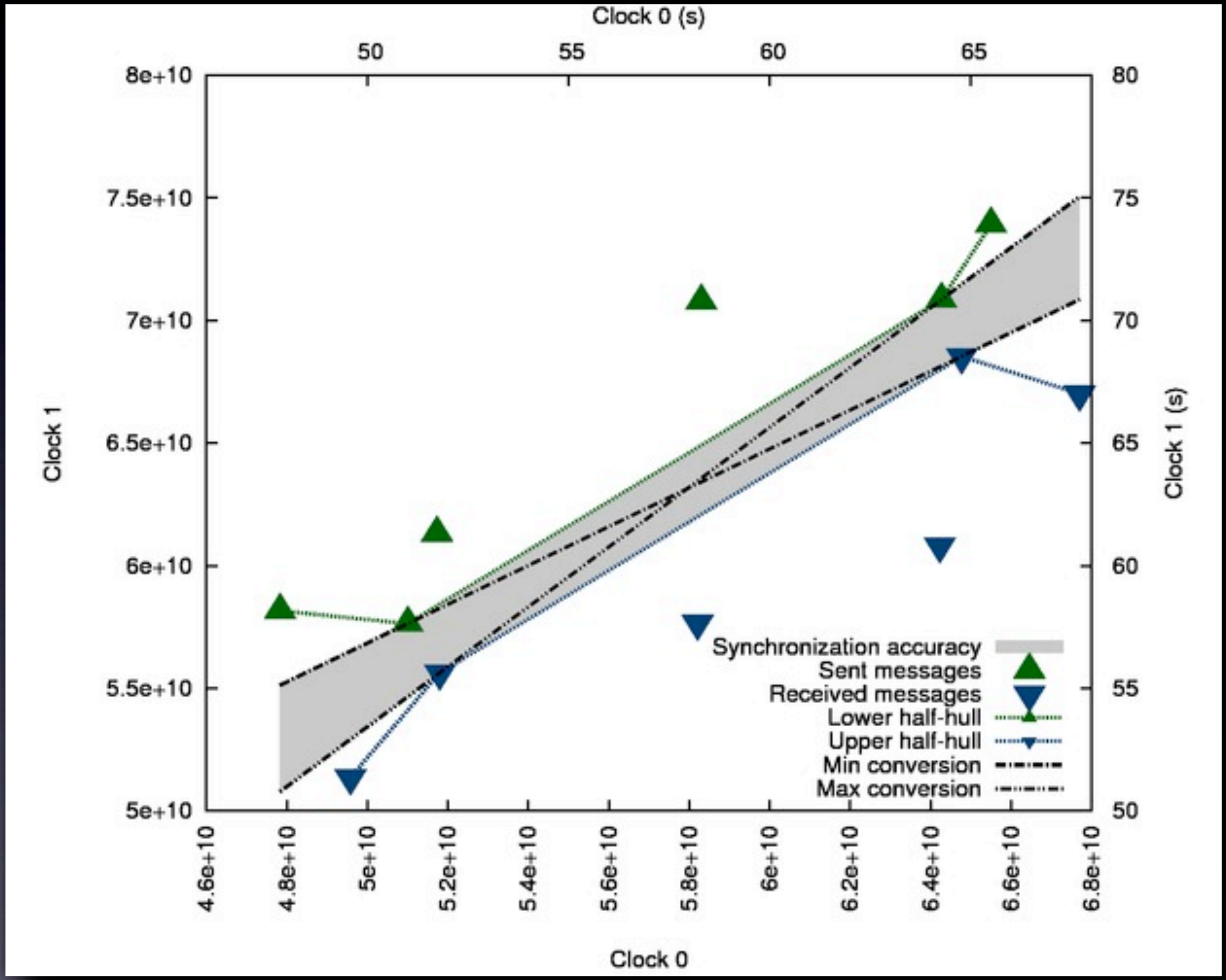
$-\infty \leq \alpha \leq \infty$

$0 \leq \beta \leq \infty$

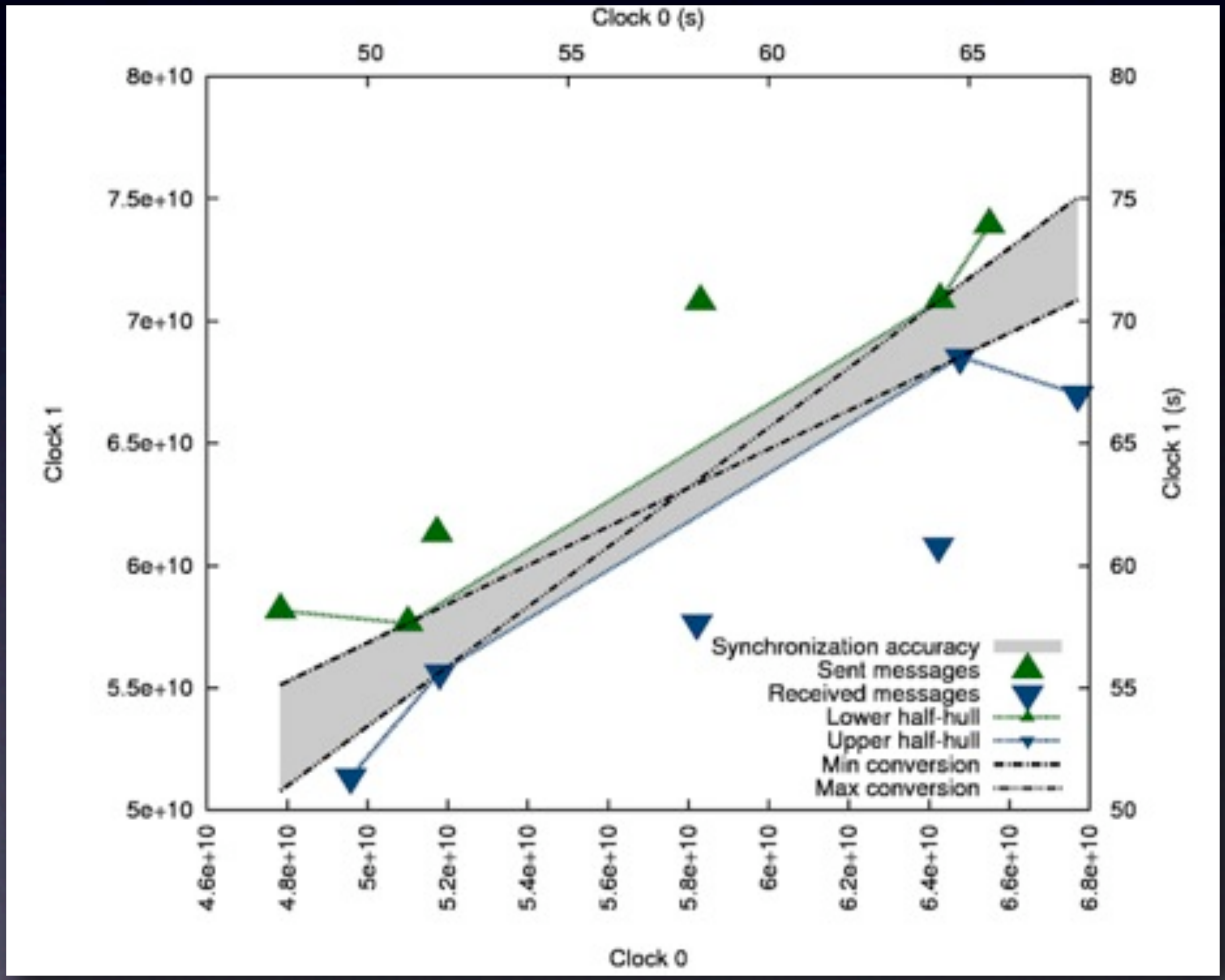
Convex Hull Method

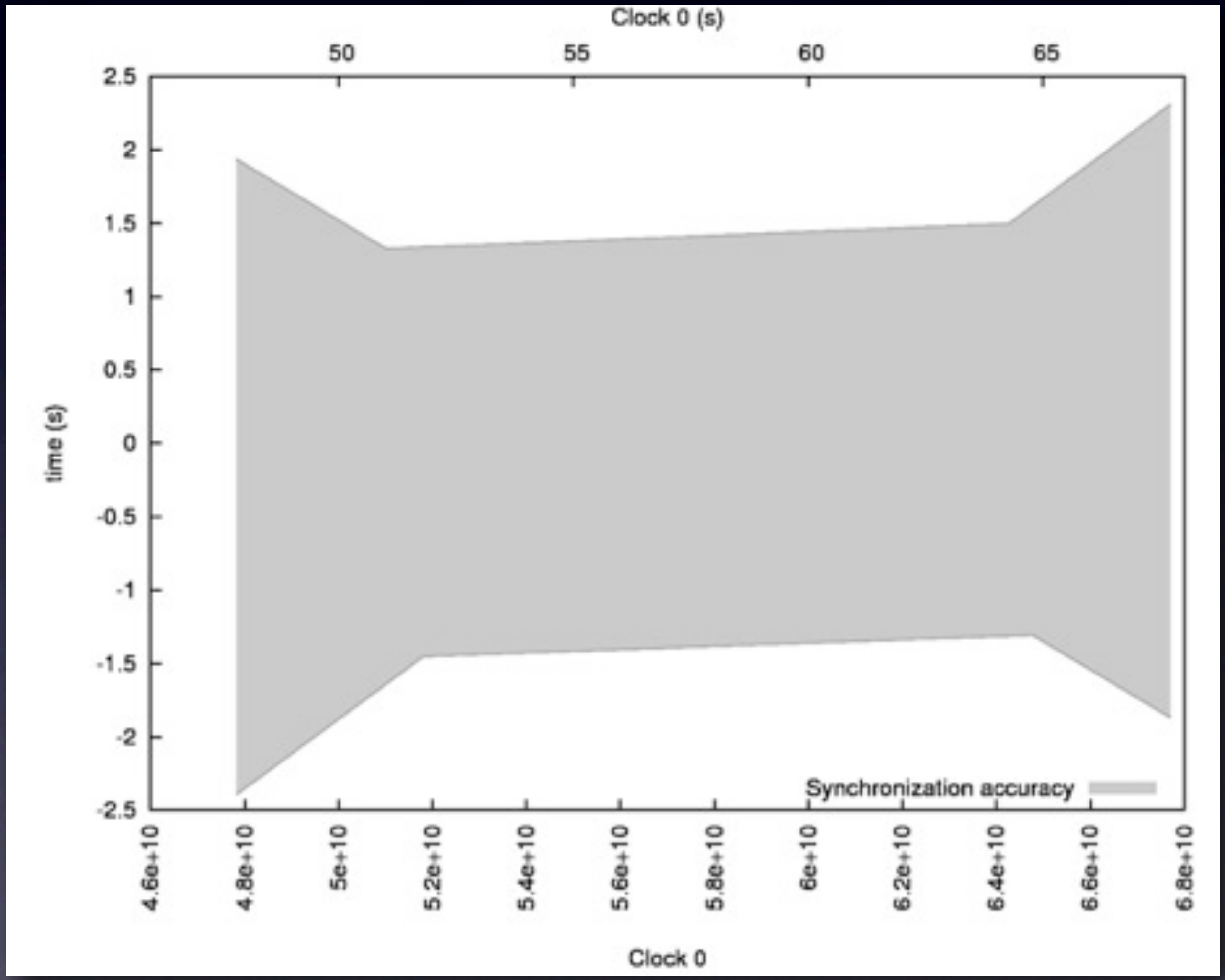


Convex Hull Method

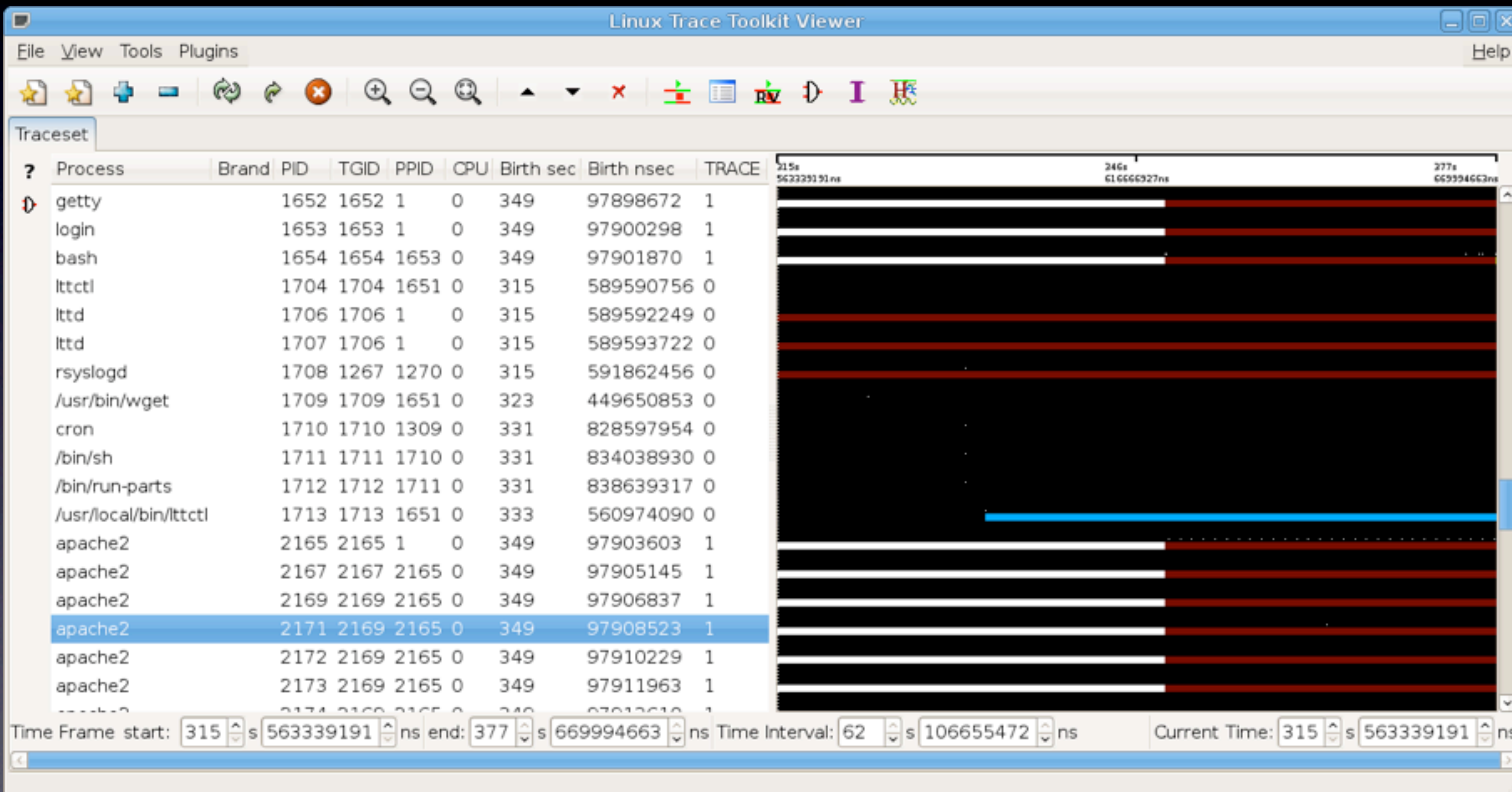


Synchronization Accuracy



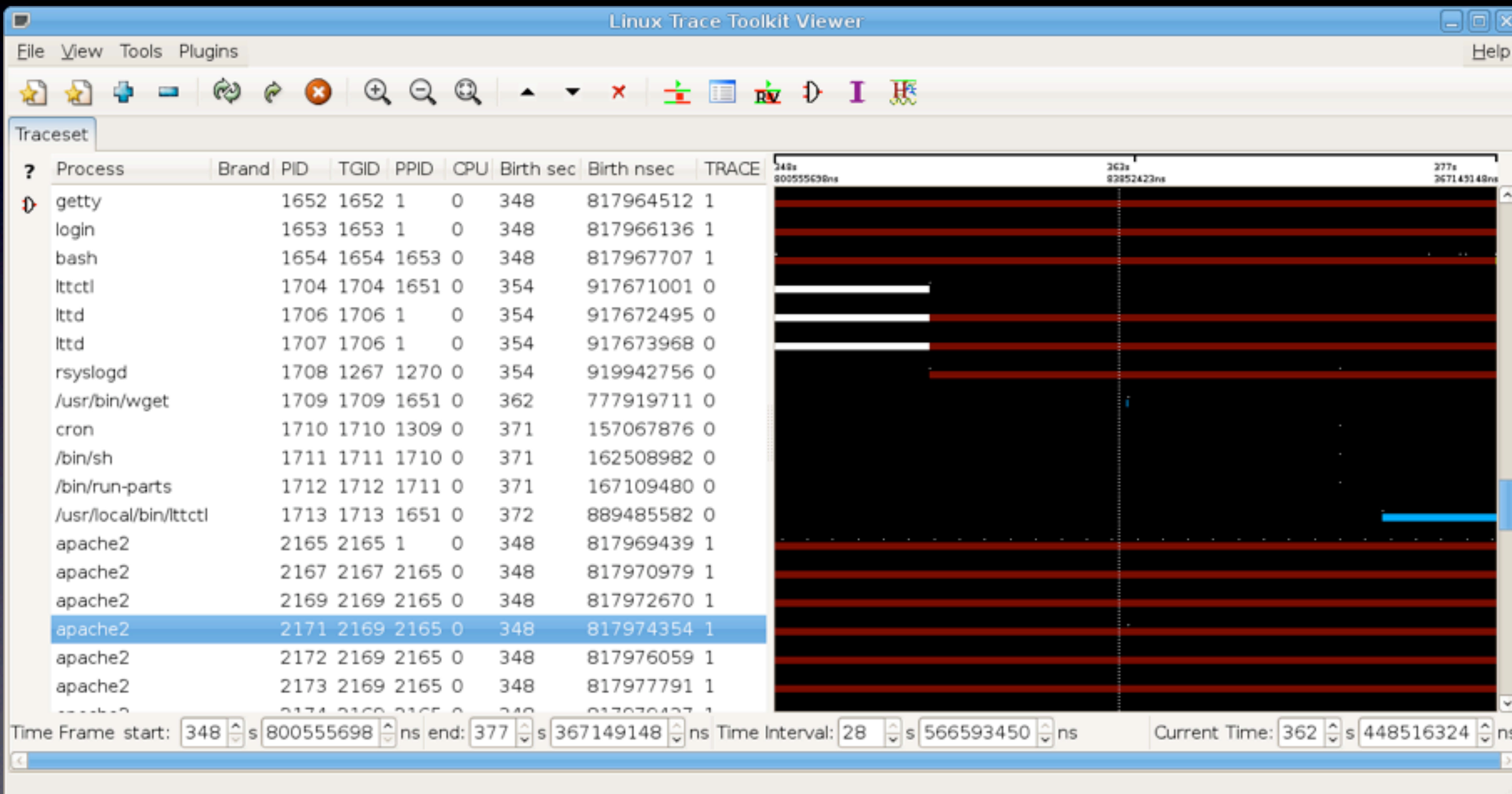


Results

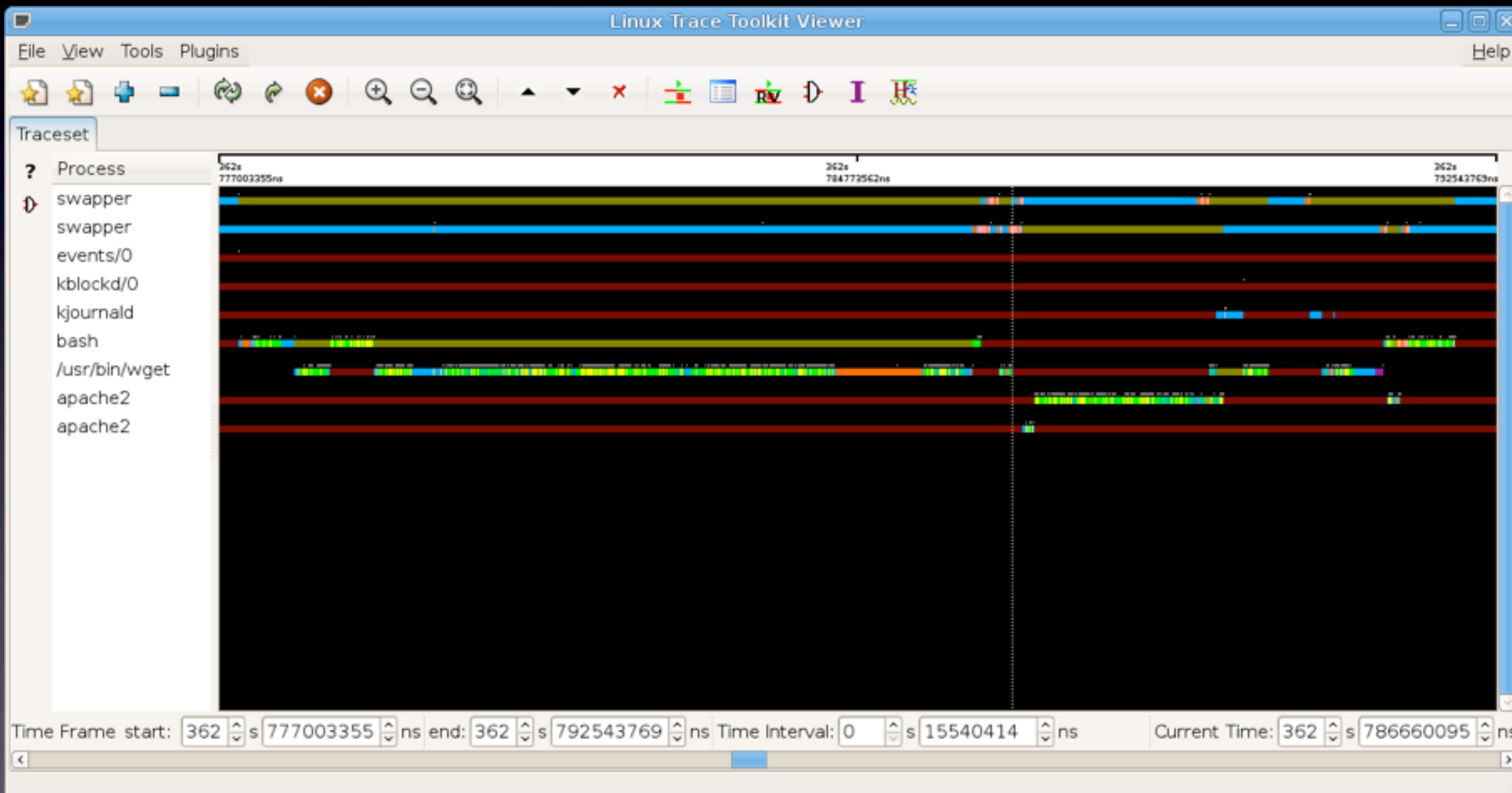


Linux Trace Toolkit Viewer
File View Tools Plugins
Help
Traceset
Process Brand PID TGID PPID CPU Birth sec Birth nsec TRACE
getty 1652 1652 1 0 349 97898672 1
login 1653 1653 1 0 349 97900298 1
bash 1654 1654 1653 0 349 97901870 1
lttctl 1704 1704 1651 0 315 589590756 0
lttd 1706 1706 1 0 315 589592249 0
lttd 1707 1706 1 0 315 589593722 0
rsyslogd 1708 1267 1270 0 315 591862456 0
/usr/bin/wget 1709 1709 1651 0 323 449650853 0
cron 1710 1710 1309 0 331 828597954 0
/bin/sh 1711 1711 1710 0 331 834038930 0
/bin/run-parts 1712 1712 1711 0 331 838639317 0
/usr/local/bin/lttctl 1713 1713 1651 0 333 560974090 0
apache2 2165 2165 1 0 349 97903603 1
apache2 2167 2167 2165 0 349 97905145 1
apache2 2169 2169 2165 0 349 97906837 1
apache2 2171 2169 2165 0 349 97908523 1
apache2 2172 2169 2165 0 349 97910229 1
apache2 2173 2169 2165 0 349 97911963 1
Time Frame start: 315 s 563339191 ns end: 377 s 669994663 ns Time Interval: 62 s 106655472 ns
Current Time: 315 s 563339191 ns

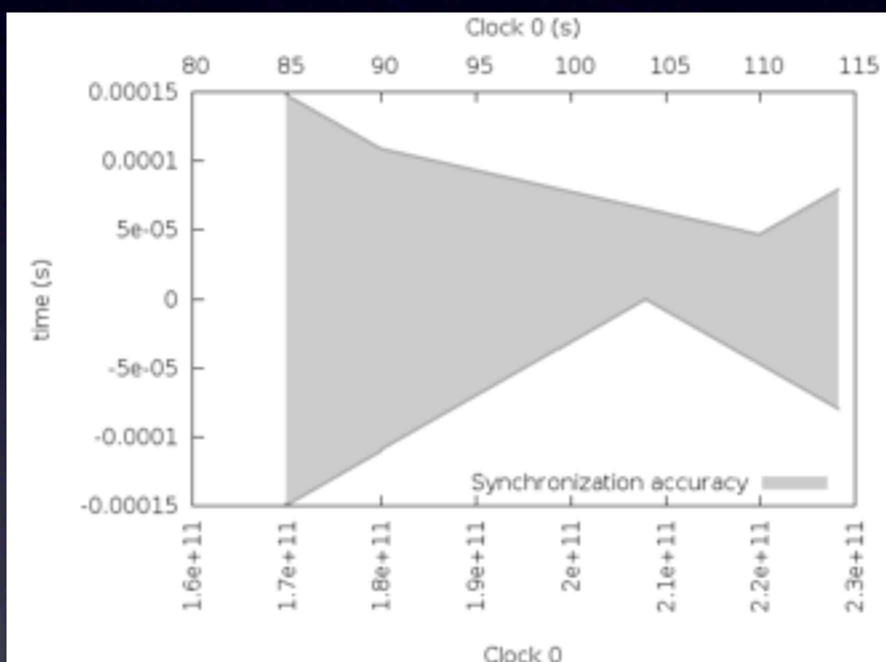
Results


Linux Trace Toolkit Viewer
File View Tools Plugins Help
Traceset
Process Brand PID TGID PPID CPU Birth sec Birth nsec TRACE
getty 1652 1652 1 0 348 817964512 1
login 1653 1653 1 0 348 817966136 1
bash 1654 1654 1653 0 348 817967707 1
lttctl 1704 1704 1651 0 354 917671001 0
lttd 1706 1706 1 0 354 917672495 0
lttd 1707 1706 1 0 354 917673968 0
rsyslogd 1708 1267 1270 0 354 919942756 0
/usr/bin/wget 1709 1709 1651 0 362 777919711 0
cron 1710 1710 1309 0 371 157067876 0
/bin/sh 1711 1711 1710 0 371 162508982 0
/bin/run-parts 1712 1712 1711 0 371 167109480 0
/usr/local/bin/lttctl 1713 1713 1651 0 372 889485582 0
apache2 2165 2165 1 0 348 817969439 1
apache2 2167 2167 2165 0 348 817970979 1
apache2 2169 2169 2165 0 348 817972670 1
apache2 2171 2169 2165 0 348 817974354 1
apache2 2172 2169 2165 0 348 817976059 1
apache2 2173 2169 2165 0 348 817977791 1
Time Frame start: 348 s 800555698 ns end: 377 s 367149148 ns Time Interval: 28 s 566593450 ns
Current Time: 362 s 448516324 ns

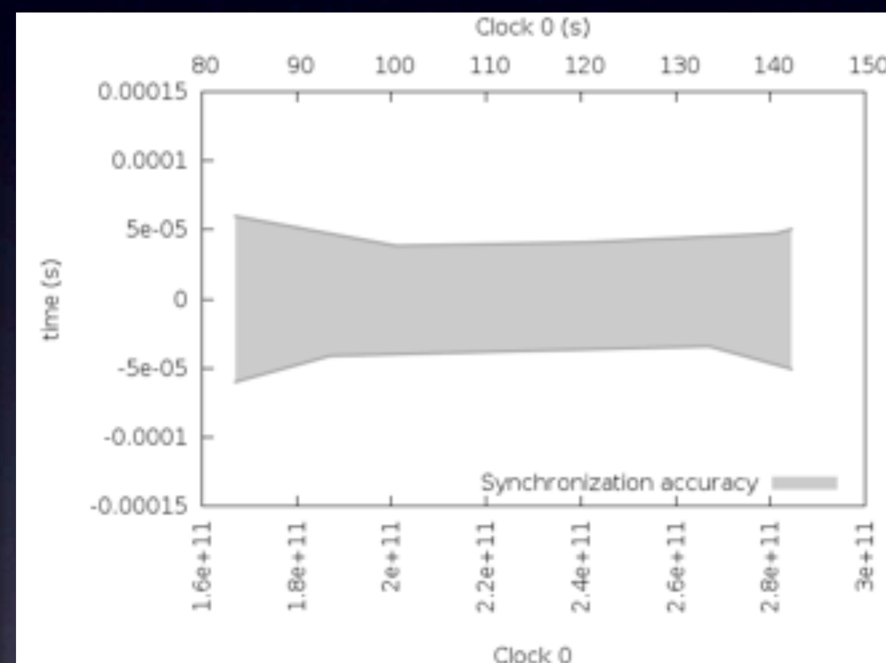
Results


Linux Trace Toolkit Viewer
File View Tools Plugins
Help
Traceset
Process
swapper
swapper
events/0
kblockd/0
kjournald
bash
/usr/bin/wget
apache2
apache2
Time Frame start: 362 s 777003355 ns end: 362 s 792543769 ns Time Interval: 0 s 15540414 ns
Current Time: 362 s 786660095 ns

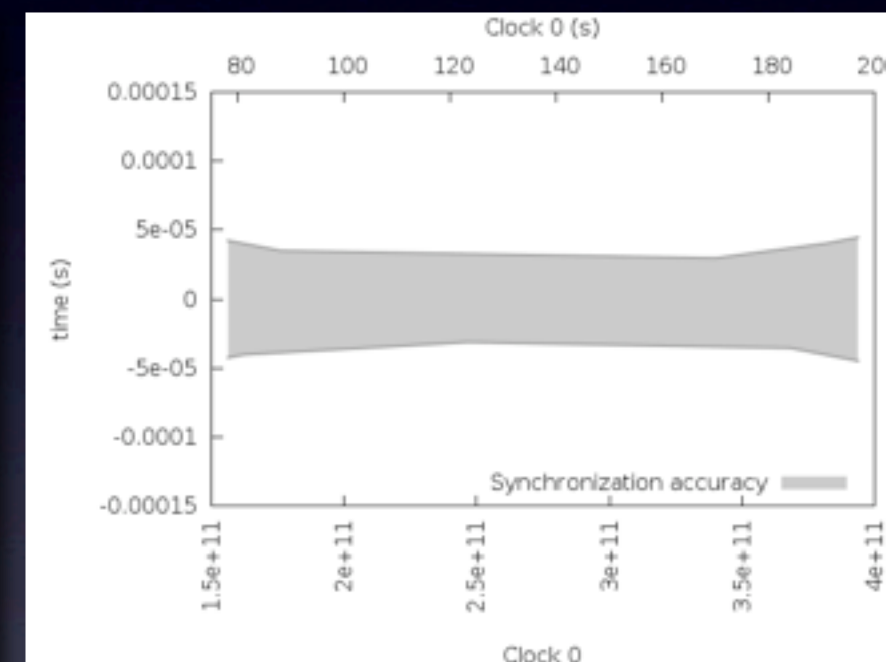
Synchronization Accuracy



30s.

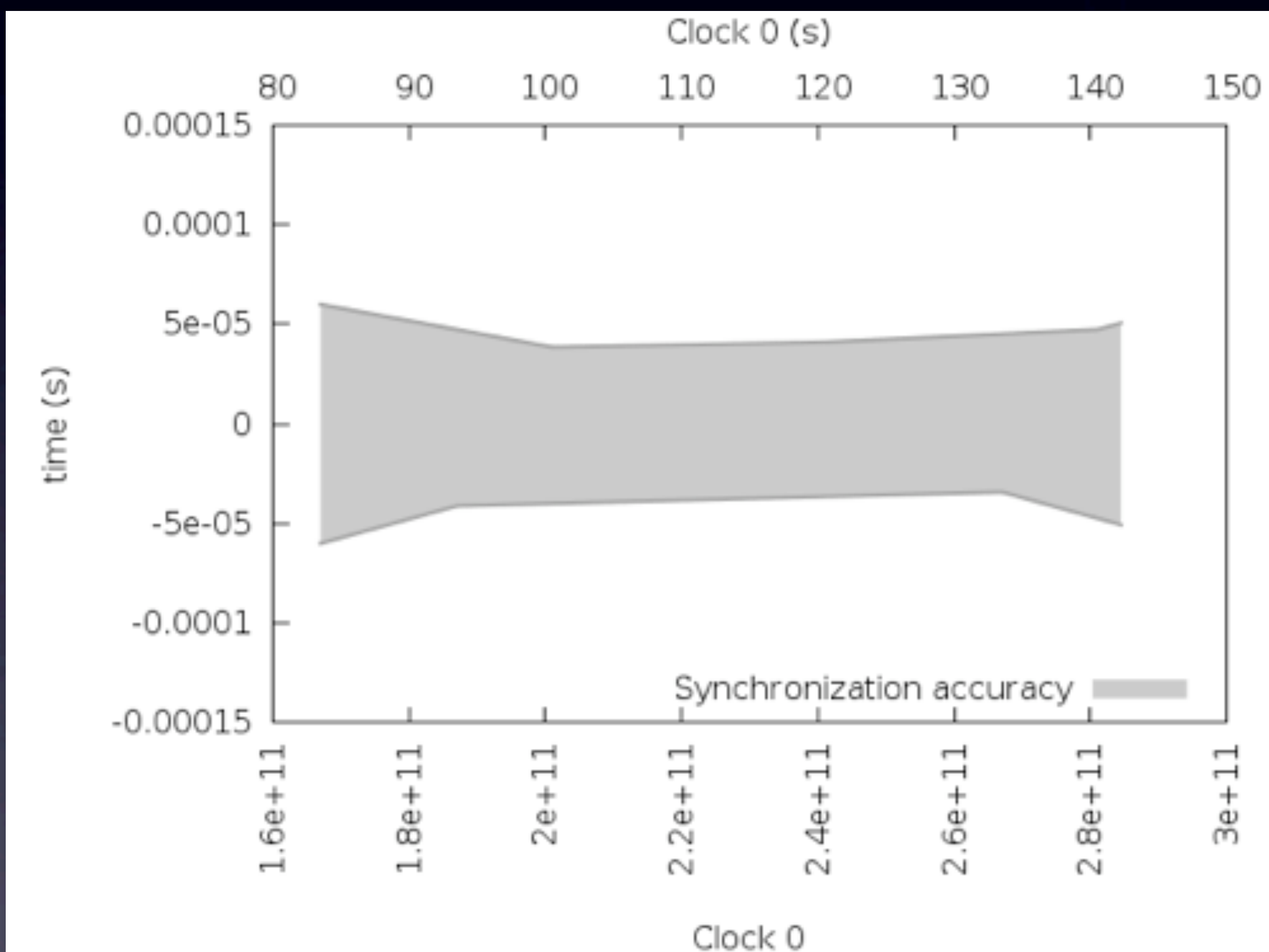


60s.

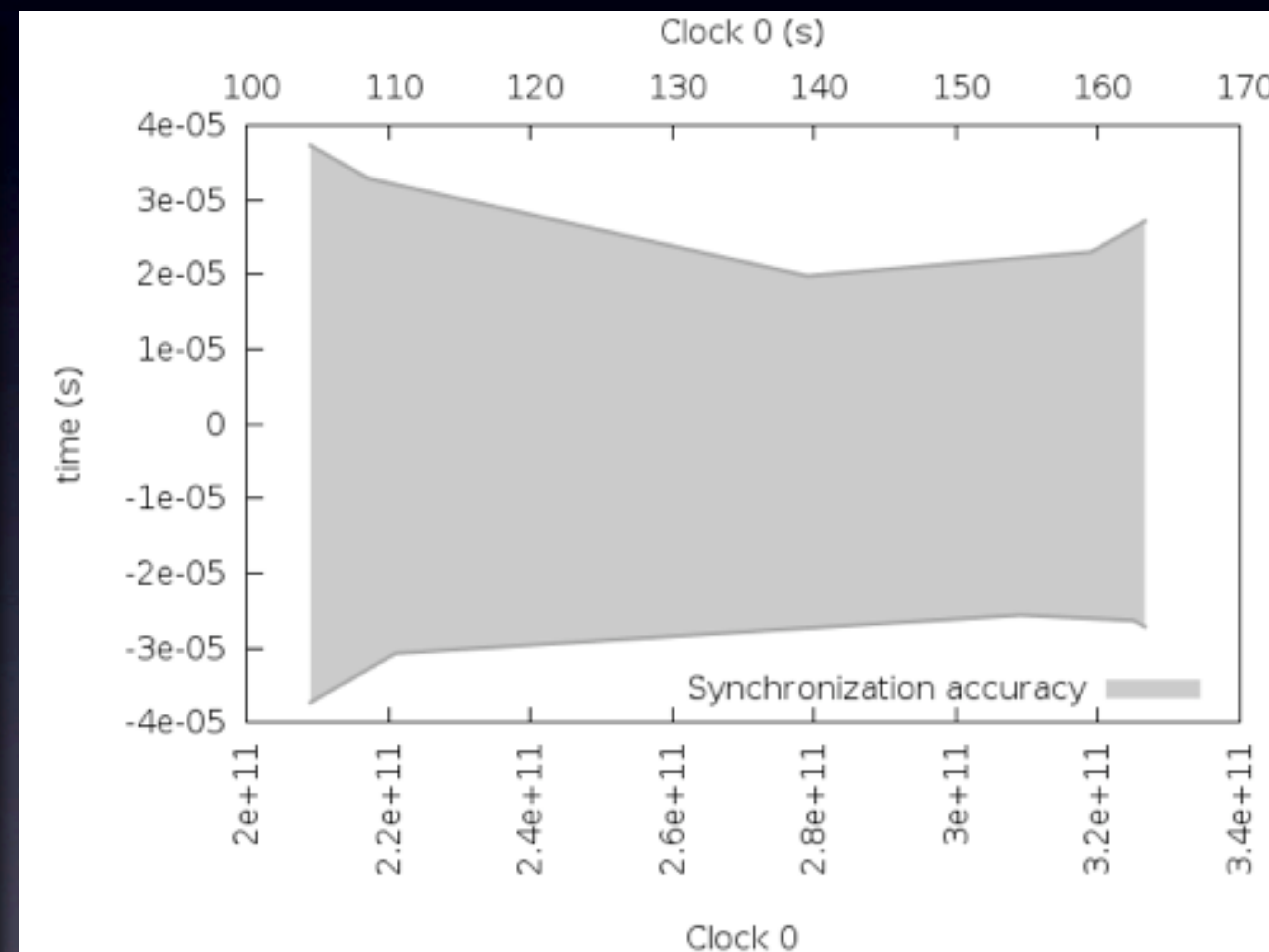


120s.

Synchronization Accuracy

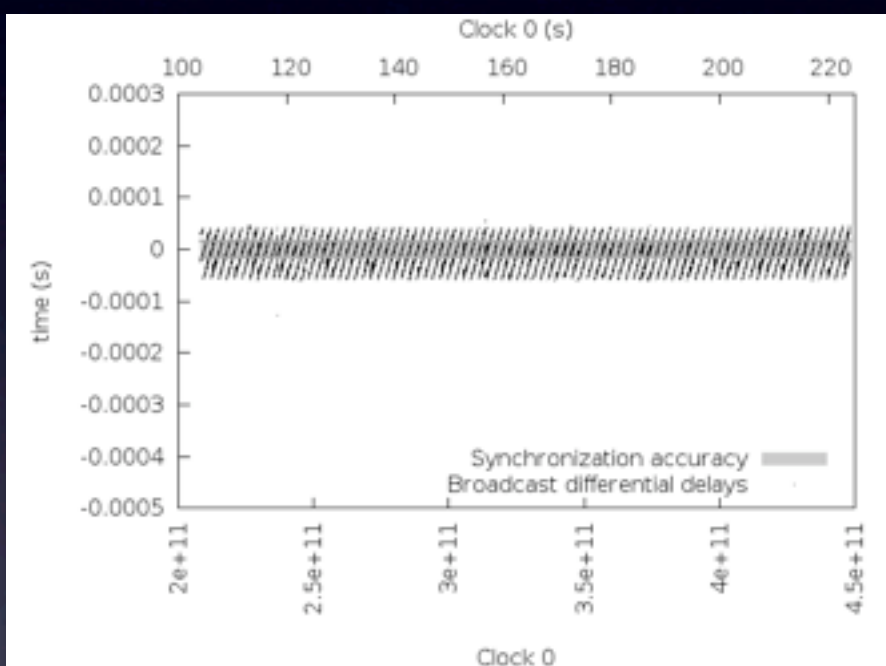


Fast Ethernet (100Mbs)

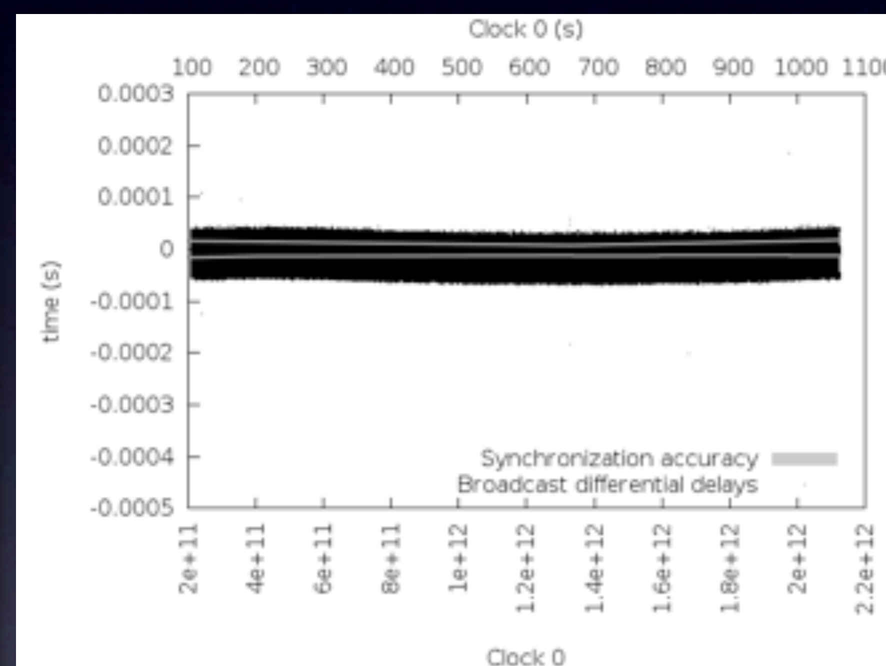


Gigabit Ethernet

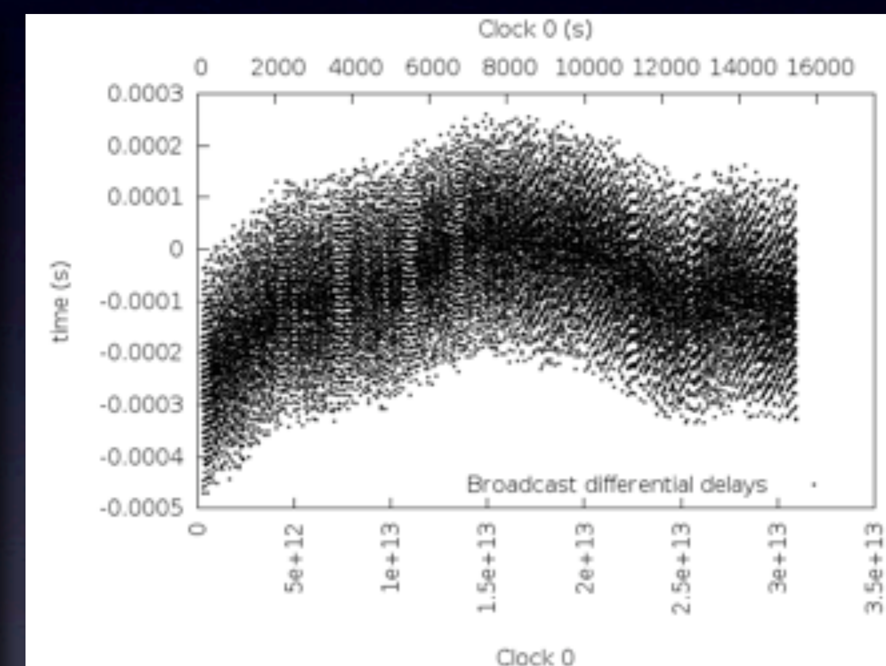
Synchronization Accuracy



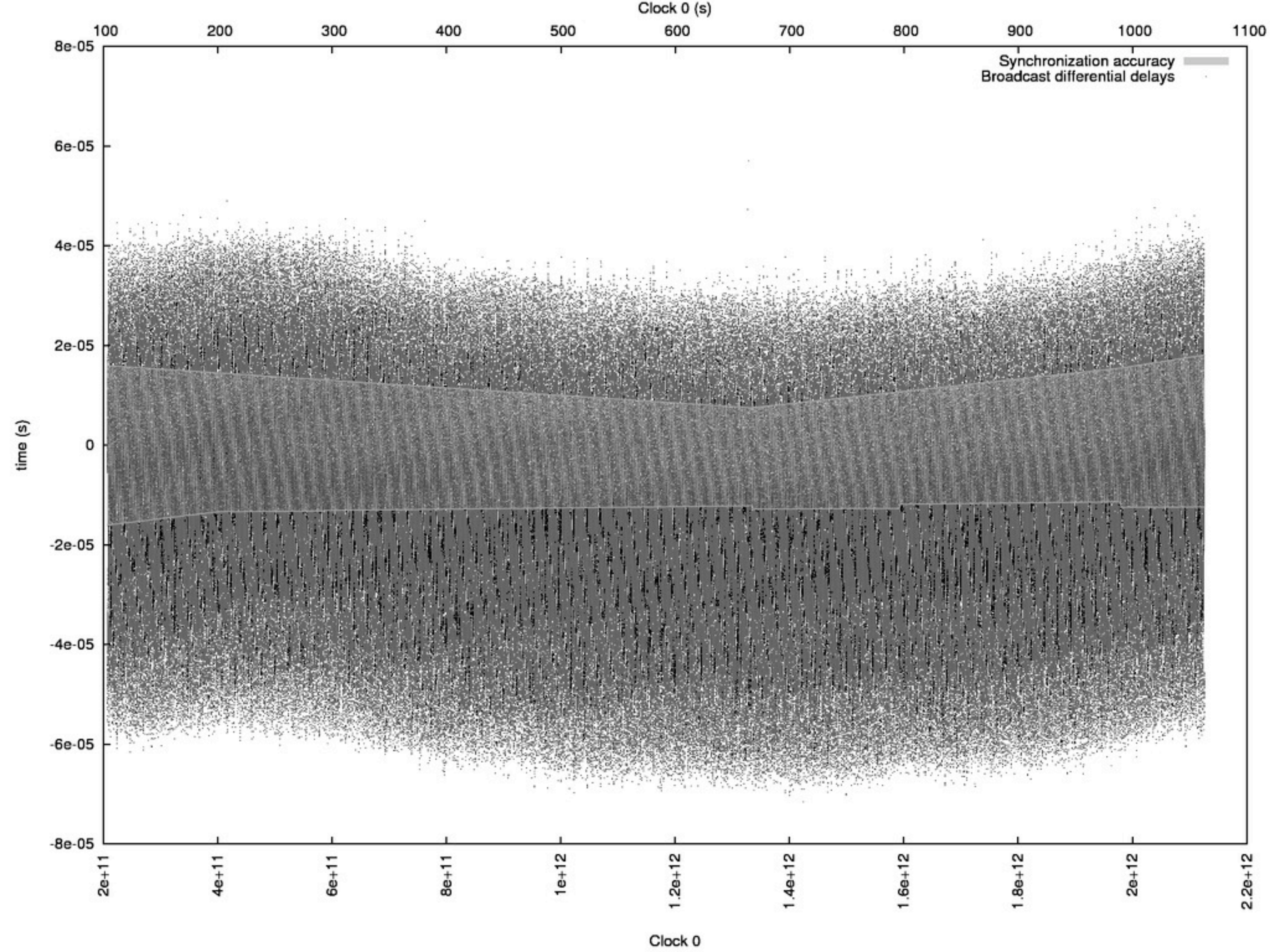
120s.



960s.
(16 minutes)



15360s.
(4:16 hours)


Clock 0 (s)
100
200
300
400
500
600
700
800
900
1000
1100
Synchronization accuracy
Broadcast differential delays
8e-05
6e-05
4e-05
2e-05
0
-2e-05
-4e-05
-6e-05
-8e-05
time (s)
2e+11
4e+11
6e+11
8e+11
1e+12
1.2e+12
1.4e+12
1.6e+12
1.8e+12
2e+12
2.2e+12
Clock 0

Statistics

```
Event count (2 traces):
   2.1-2nodes-1024/noeud1: 1140075
   2.1-2nodes-1024/noeud2: 1140285
   total events: 2280360
LTTV processing stats:
   received frames: 922900
   received frames that are IP: 922888
   received and processed packets that are TCP: 553952
   received and processed packets that are UDP: 245766
   sent packets that are TCP: 553813
TCP matching stats:
   Message traffic:
        0 - 1  : sent 215138     received 215076
Broadcast matching stats:
   total broadcasts datagrams received: 245762
   total broadcast groups for which all receptions were identified:
      122881
```

Statistics

```
Linear regression analysis stats:
     Individual synchronization factors:
           0 - 1  : a0=  3.81325e+08 a1= 1 + 6.16834e-06 accuracy 23092.8
           1 - 0  : a0= -3.81251e+08 a1= 1 - 6.18234e-06 accuracy 23876.0

Convex hull analysis stats:
     out of order packets dropped from analysis: 0
     Number of points in convex hulls:
           0 - 1  : lower half-hull 16    upper half-hull 14
     Individual synchronization factors:
           0 - 1  : Middle
                    a0=  3.81288e+08 a1= 1 + 6.17755e-06 accuracy 5.54264e-07
                    a0:  3.81197e+08 to  3.81378e+08 (delta= 181440)
                    a1: 1 +5.90042e-06 to +6.45468e-06 (delta= 5.54264e-07)
```

Statistics

```
Linear regression analysis stats:
     Individual synchronization factors:
           0 - 1  : a0=  3.81325e+08 a1= 1 + 6.16834e-06 accuracy 23092.8
           1 - 0  : a0= -3.81251e+08 a1= 1 - 6.18234e-06 accuracy 23876.0

Convex hull analysis stats:
     out of order packets dropped from analysis: 0
     Number of points in convex hulls:
           0 - 1  : lower half-hull 16    upper half-hull 14
     Individual synchronization factors:
           0 - 1  : Middle
                    a0=  3.81288e+08 a1= 1 + 6.17755e-06 accuracy 5.54264e-07
                    a0:  3.81197e+08 to  3.81378e+08 (delta= 181440)
                    a1: 1 +5.90042e-06 to +6.45468e-06 (delta= 5.54264e-07)
```

Statistics

```
Synchronization evaluation analysis stats:
    sum of broadcast differential delays: 1.71
    average broadcast differential delay: 1.39159e-05
    Individual evaluation:
       Trace pair  Inversions       Too fast         Total
         0 - 1                      47842 (23%)      215138
         1 - 0                      0 (0%)           215076
         total                      47842 (12%)      430214

Synchronization evaluation analysis stats:
    sum of broadcast differential delays: 1.11123
    average broadcast differential delay: 9.04313e-06
    Individual evaluation:
       Trace pair  Inversions       Too fast         Total
         0 - 1     0 (0%)           18243 (9%)       215138
         1 - 0     0 (0%)           6951 (4%)        215076
         total     0 (0%)           25194 (6%)       430214
```

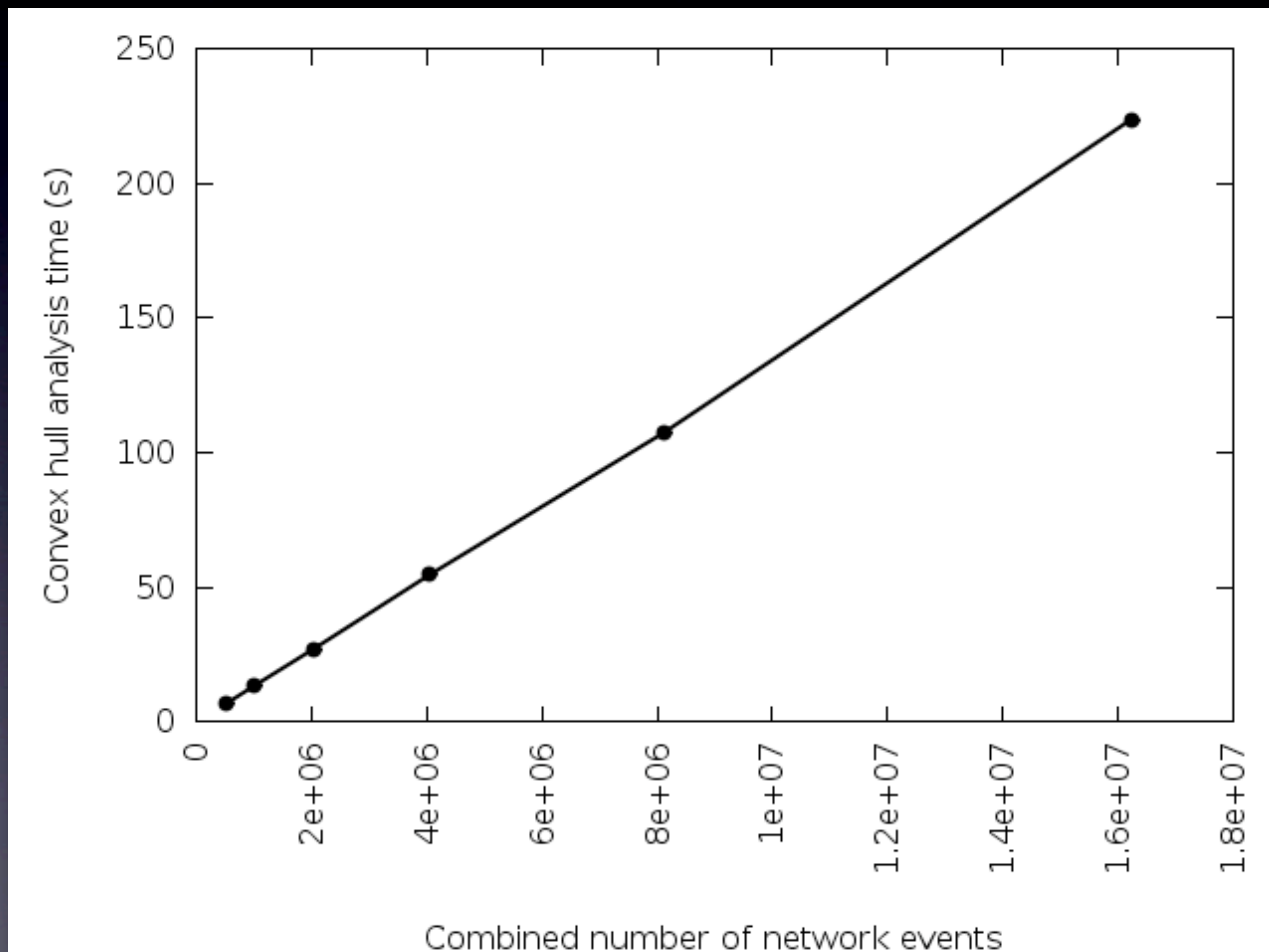
Statistics

```
Synchronization evaluation analysis stats:
    sum of broadcast differential delays: 1.71
    average broadcast differential delay: 1.39159e-05
    Individual evaluation:
       Trace pair  Inversions        Too fast          Total
         0 - 1     156 (1%)          47842 (23%)       215138
         1 - 0     0 (0%)            0 (0%)            215076
         total     156 (1%)          47842 (12%)       430214

Synchronization evaluation analysis stats:
    sum of broadcast differential delays: 1.11123
    average broadcast differential delay: 9.04313e-06
    Individual evaluation:
       Trace pair  Inversions        Too fast          Total
         0 - 1     0 (0%)            18243 (9%)        215138
         1 - 0     0 (0%)            6951 (4%)         215076
         total     0 (0%)            25194 (6%)        430214
```

Statistics

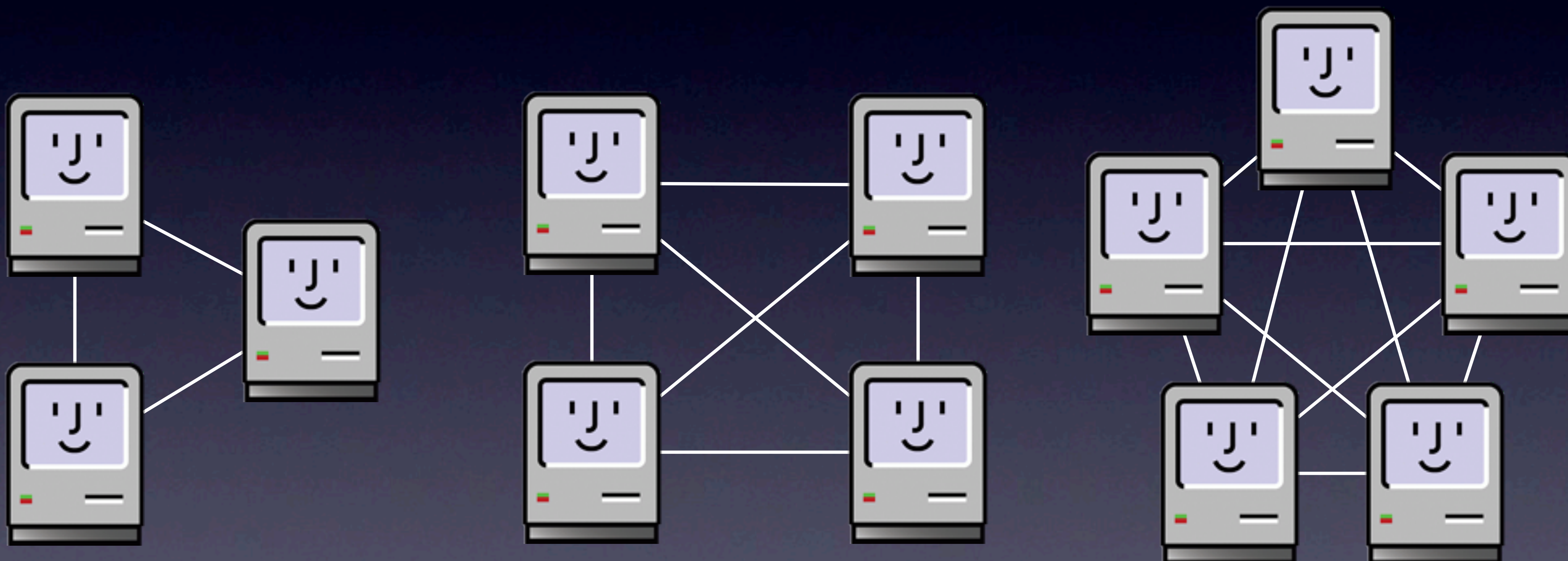
```
Synchronization evaluation analysis stats:
    sum of broadcast differential delays: 1.71
    average broadcast differential delay: 1.39159e-05
    Individual evaluation:
       Trace pair  Inversions       Too fast          Total
         0 - 1     156 (1%)         47842 (23%)       215138
         1 - 0     0 (0%)           0 (0%)            215076
         total     156 (1%)         47842 (12%)       430214

Synchronization evaluation analysis stats:
    sum of broadcast differential delays: 1.11123
    average broadcast differential delay: 9.04313e-06
    Individual evaluation:
       Trace pair  Inversions       Too fast          Total
         0 - 1     0 (0%)           18243 (9%)        215138
         1 - 0     0 (0%)           6951 (4%)         215076
         total     0 (0%)           25194 (6%)        430214
```

Runtime Performance



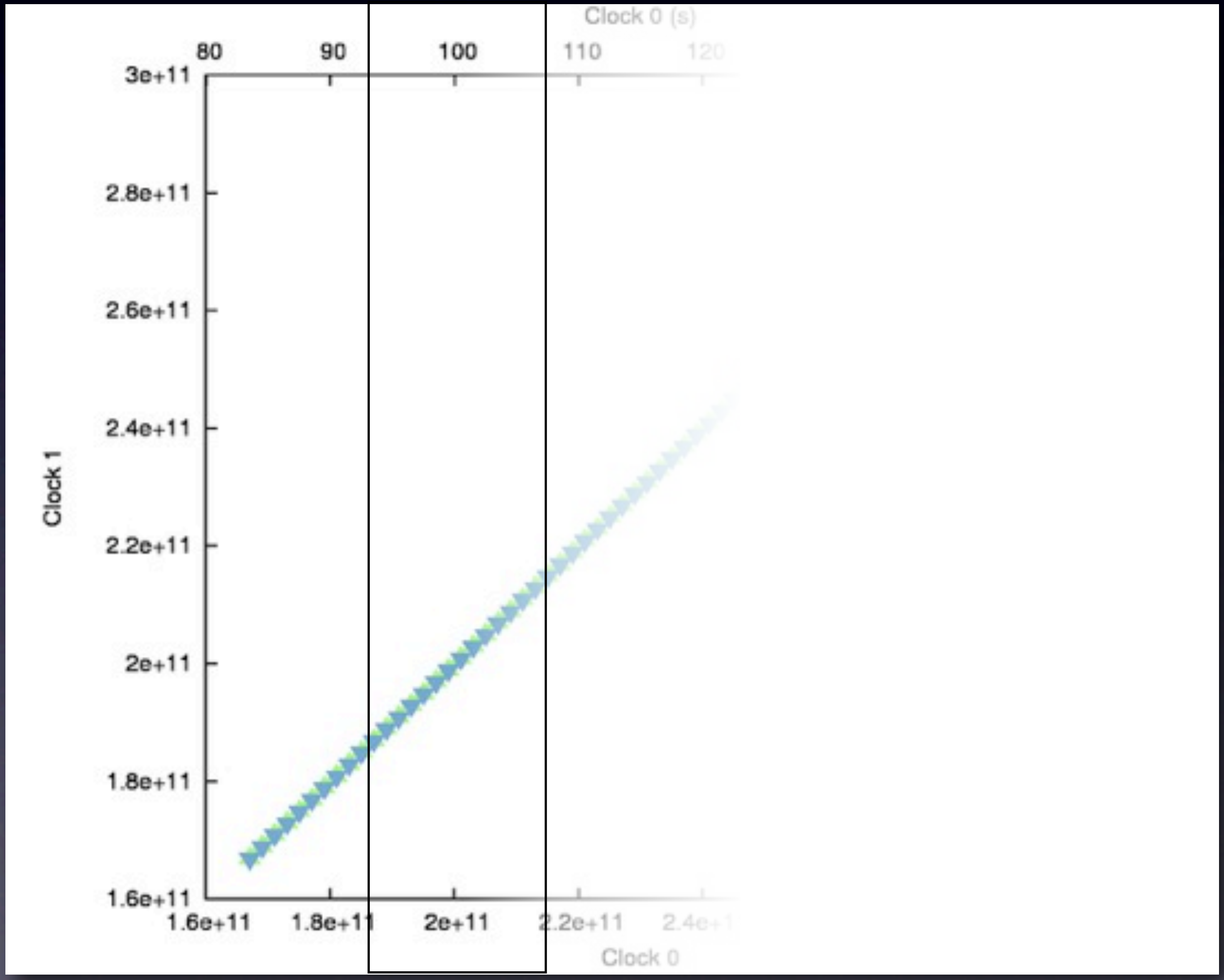
Future Developments

Synchronization across more than two nodes



Future Developments

Streaming trace synchronization

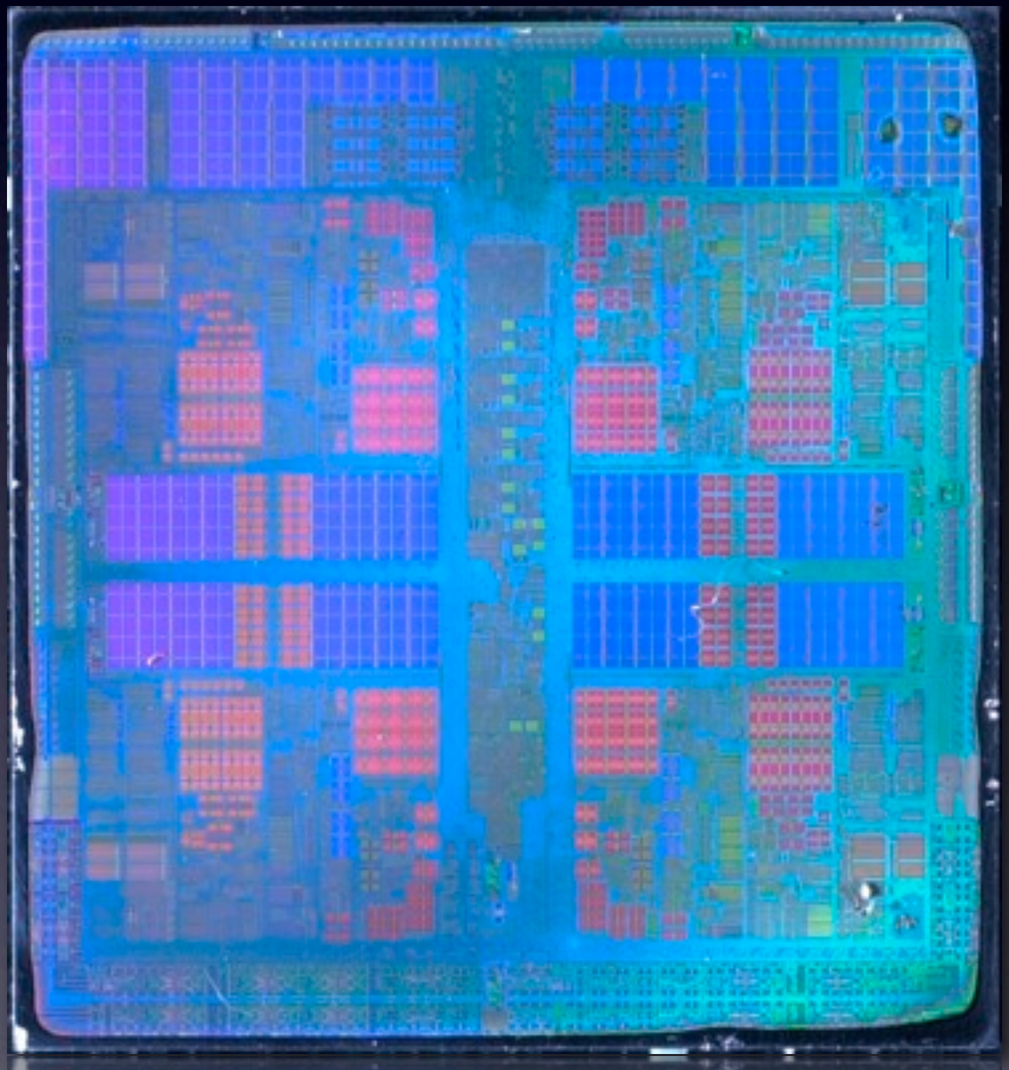


Future Developments

Tracing on future multi-core architectures

Future Developments

Tracing on future multi-core architectures

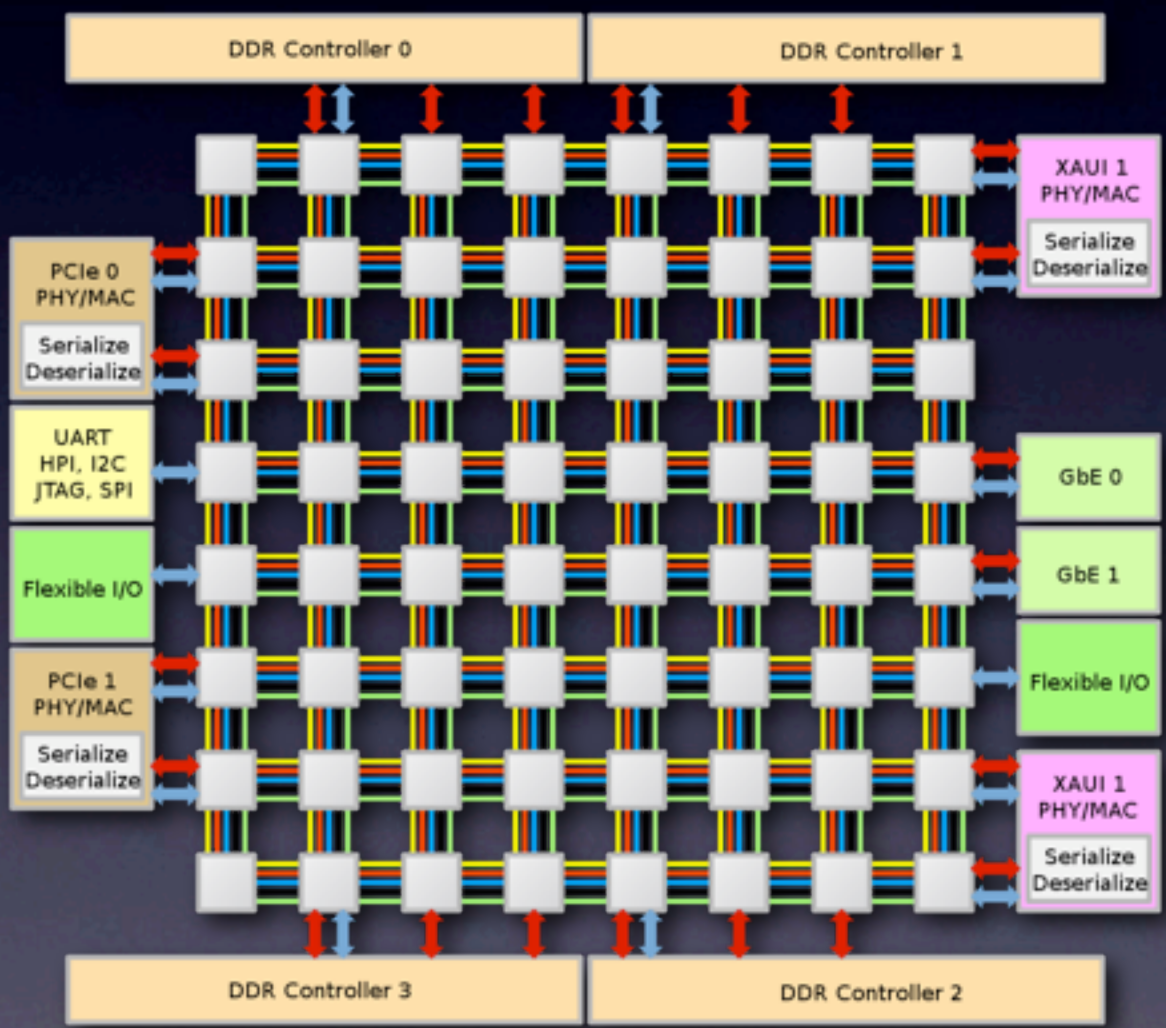


AMD K10 (Opteron, Phenom)

- 2 to 6 cores
- Three-level cache
- Integrated memory controller
- HyperTransport bus

Future Developments

Tracing on future multi-core architectures

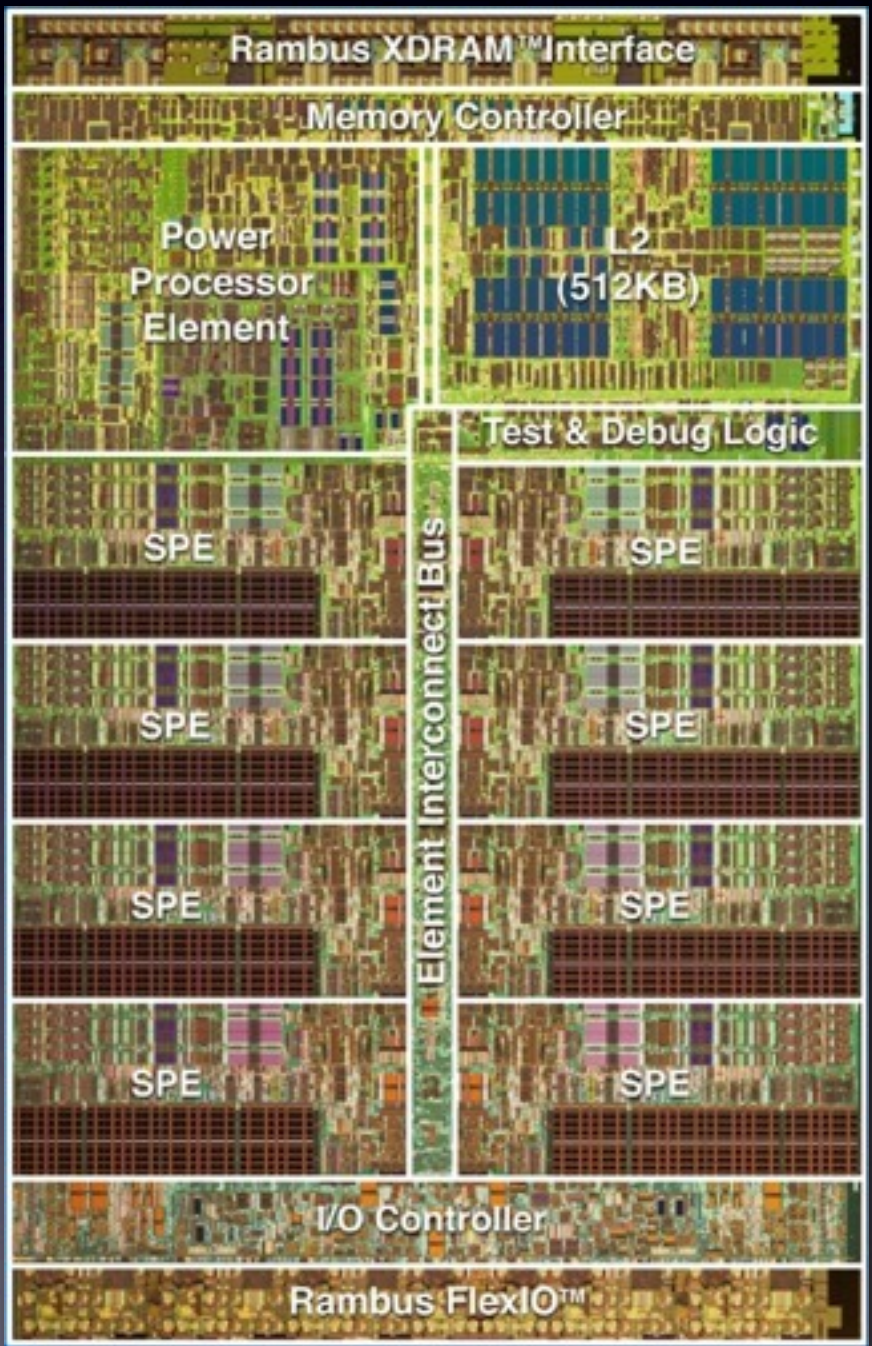


Tilera TILE64

- 64 cores
- Integrated memory controllers
- On-chip five lane "mesh" network

Future Developments

Tracing on future multi-core architectures



STI Cell Broadband Engine

- 9 heterogenous cores
- Integrated memory controller
- Multi-lane ring bus

Future Developments

Tracing on future multi-core architectures

Intel SCC

- 48 cores
- Integrated memory controllers
- Independant voltage and frequency “islands”

Future Developments

-core architectures

el SCC

48 cores

Integrated memory controllers

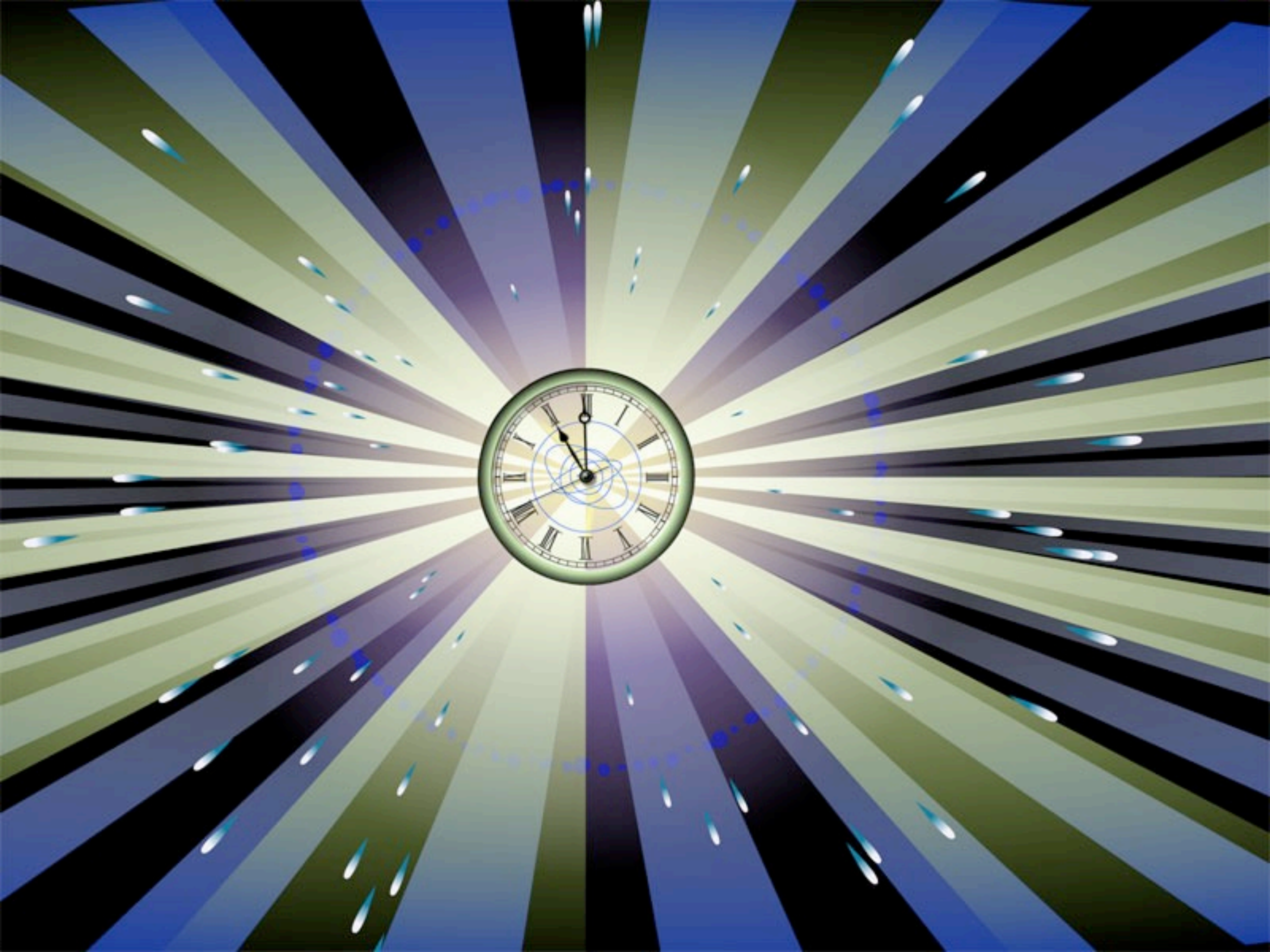
Independant voltage and frequency “islands”

Future Developments

Tracing on future multi-core architectures

Intel SCC

- 48 cores
- Integrated memory controllers
- Independant voltage and frequency "islands"



References

Linear regression and convex hull synchronization algorithms

- Duda, A., Harrus, G., Haddad, Y., and Bernard, G.: Estimating global time in distributed systems, Proc. 7th Int. Conf. on Distributed Computing Systems, Berlin, volume 18, 1987
- Ashton, P.: Algorithms for Off-line Clock Synchronisation, University of Canterbury, December 1995

Streaming trace synchronization

- Sirdey, R., and Maurice, F.: A linear programming approach to highly precise clock synchronization over a packet network, 4OR: A Quarterly Journal of Operations Research 6(4), volume 6, Springer, 393-401, 2008

Systems of more than 2 nodes

- Jezequel, J.M., and Jard, C.: Building a global clock for observing computations in distributed memory parallel computers, Concurrency: Practice and Experience 8(1), volume 8, John Wiley & Sons, Ltd Chichester, 1996
- Scheuermann, B., Kiess, W., Roos, M., Jarre, F., and Mauve, M.: On the Time Synchronization of Distributed Log Files in Networks With Local Broadcast Media, Networking, IEEE/ACM Transactions on 17(2), volume 17, 431-444, April 2009

Multi-core architectures

http://www.amd.com/us/products/desktop/processors/phenom/Pages/AMD-phenom-processor-X4-X3-at-home.aspx
http://www.tilera.com/products/TILE64.php
http://www.ibm.com/developerworks/power/cell/docs_documentation.html
http://www.intel.com/pressroom/archive/releases/20091202comp_sm.htm

State of the Art Meeting

Multi-level Multi-core Distributed Trace Synchronization

Benjamin Poirier