Node Health Checker

Scaling Improvements

Automatic Dump and Reboot

Kent Thomson
Cray, Inc

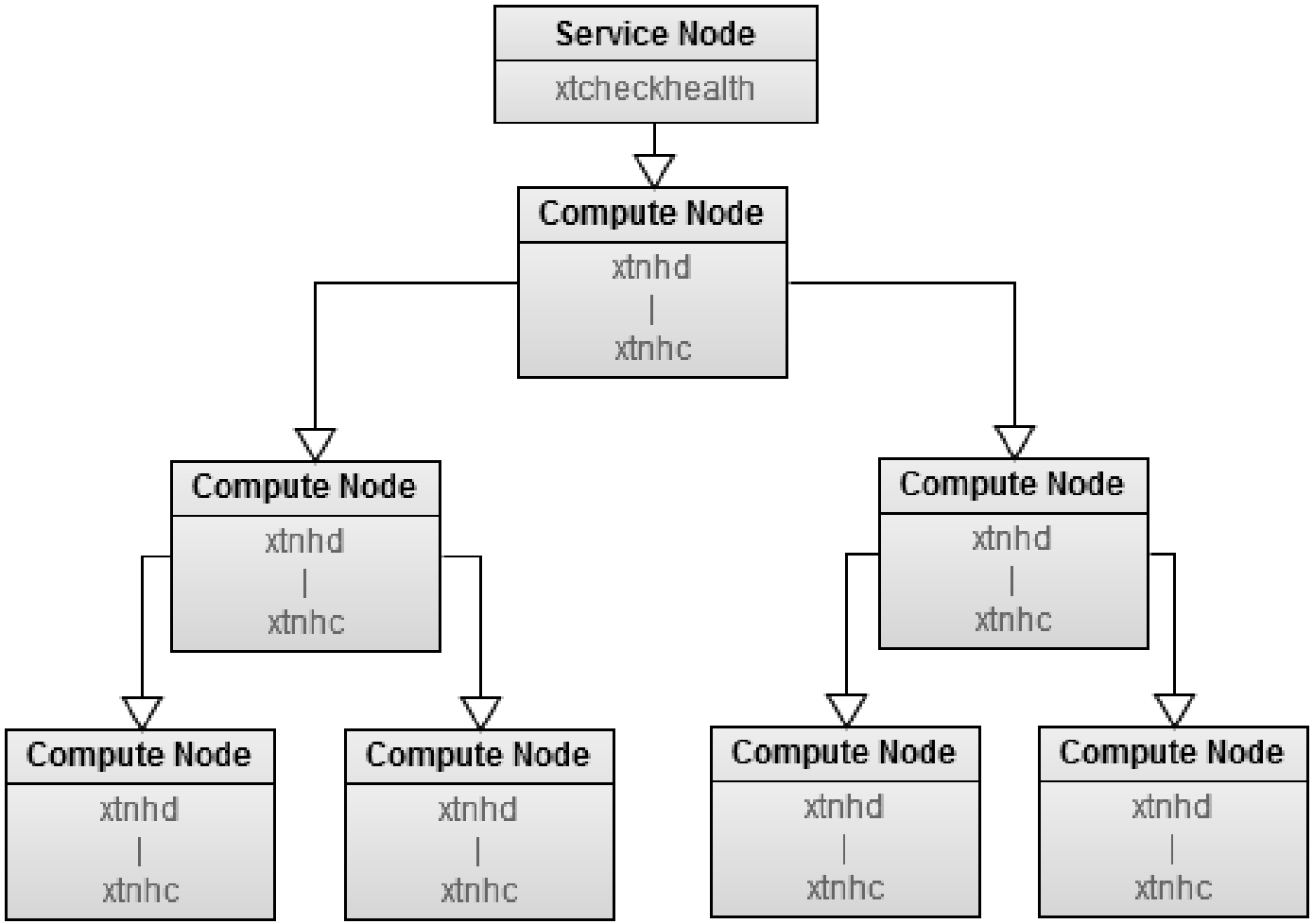
Topics

- **Node Health Checker (NHC) Overview**
  - Components
  - Architecture
  - Normal Mode vs Suspect Mode
- **NHC Scaling Improvements**
  - Initial Investigation
  - Linear Scaling
  - Scaling Fix
  - New Performance
- **Automated Dump and Reboot**
  - Design Goals
  - Dumpd SMW Daemon
  - Use Cases

NHC Overview

- **Service Node Components**
  - xtcleanup_after
  - Xtcheckhealth
  - NHC config file
- **Compute Node Components**
  - xtnhc
  - xtnhd
- **Normal Mode**
  - Initial testing of node
  - Tests run once
- **Suspect Mode**
  - Longer (35 minutes) testing of node
  - Tests restarted on failure

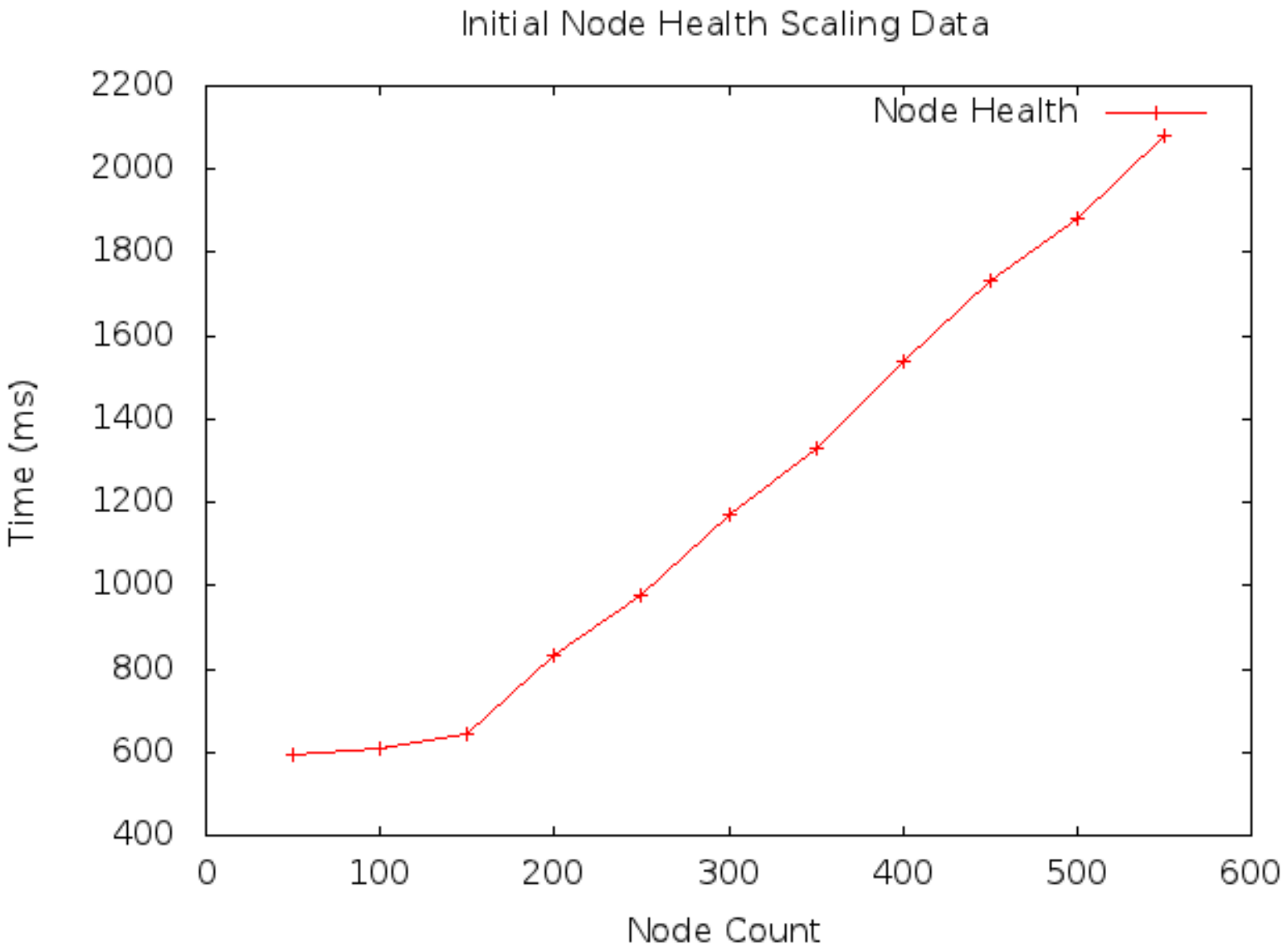
NHC Fanout Tree



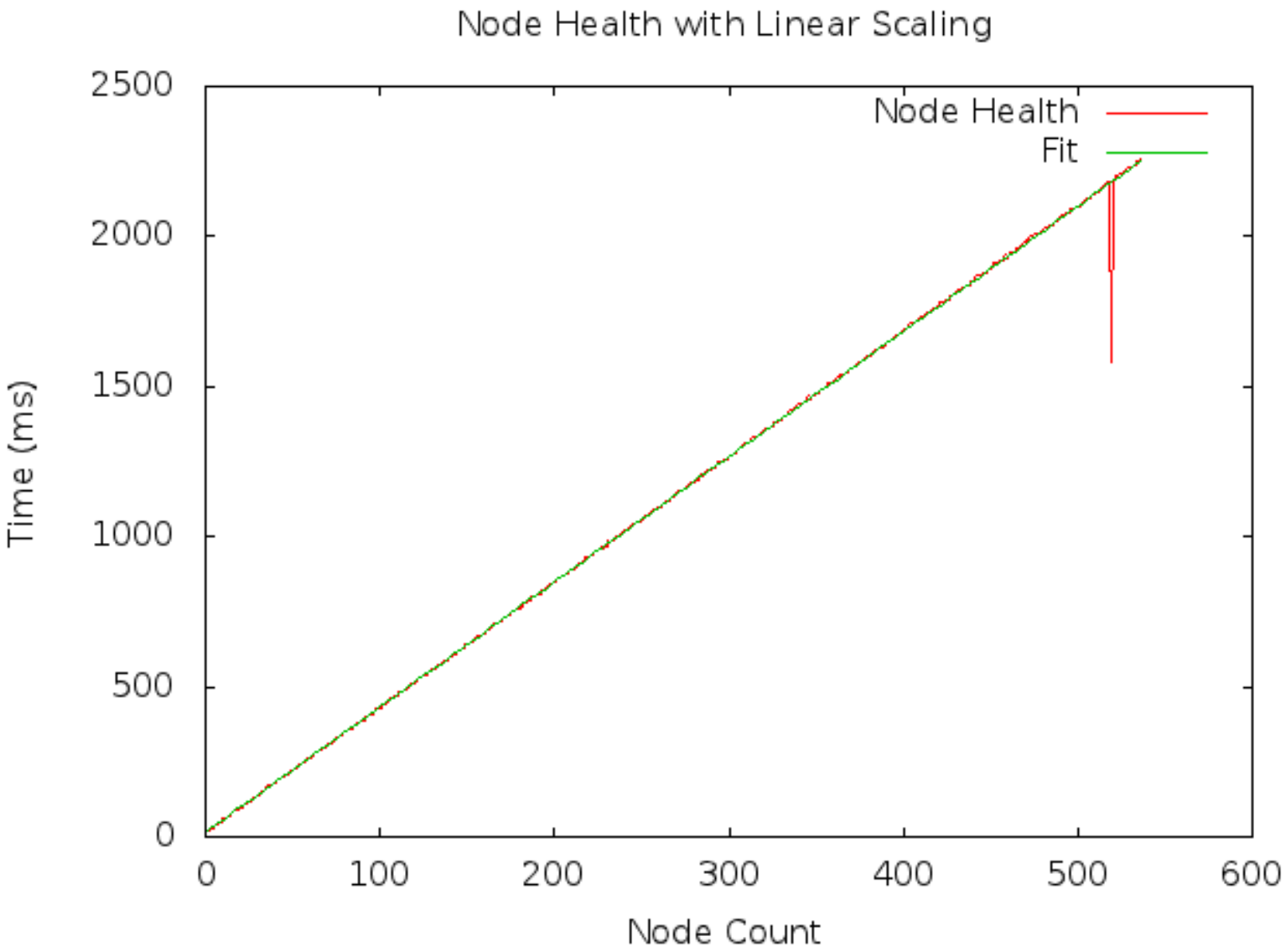
Data Collection

- Run NHC across various node counts
- Collect time to run for each node count

Initial Scaling Data



More Formal Scaling Data



Scaling Equations

- **All equations**
  - t is time in milliseconds
  - n is node count
- **Linear scaling**

$$t = 3.19n + 259$$

- **Simplifies at large node counts to**

$$t \approx 3.19n$$

- **Table of projected values**

Node Count	NHC Run Time (s)
1000	3.4
10000	32.2
20000	64.1

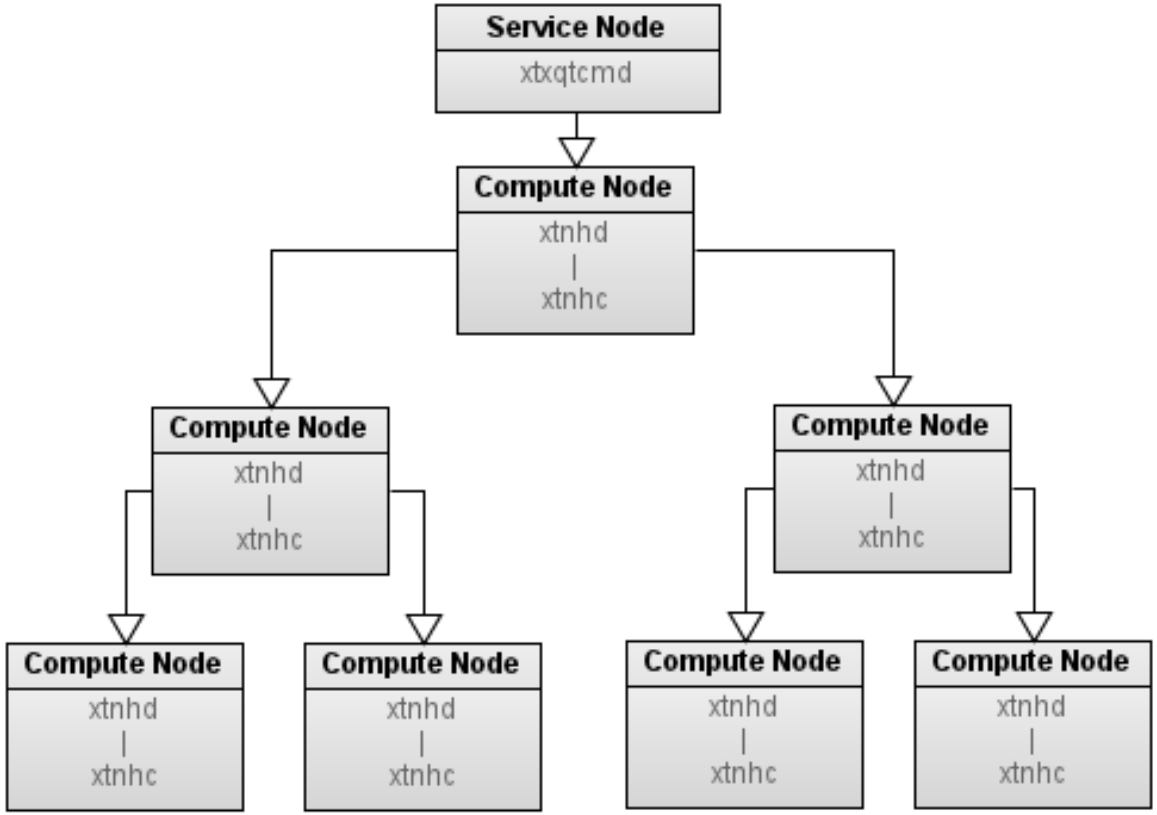
Curve Fitting and Gnuplot

- **Curve Fitting**
  - Deriving equation that closely matches experimental data
  - Can be used to identify scaling order
  - Useful for extrapolating run times out to large node counts

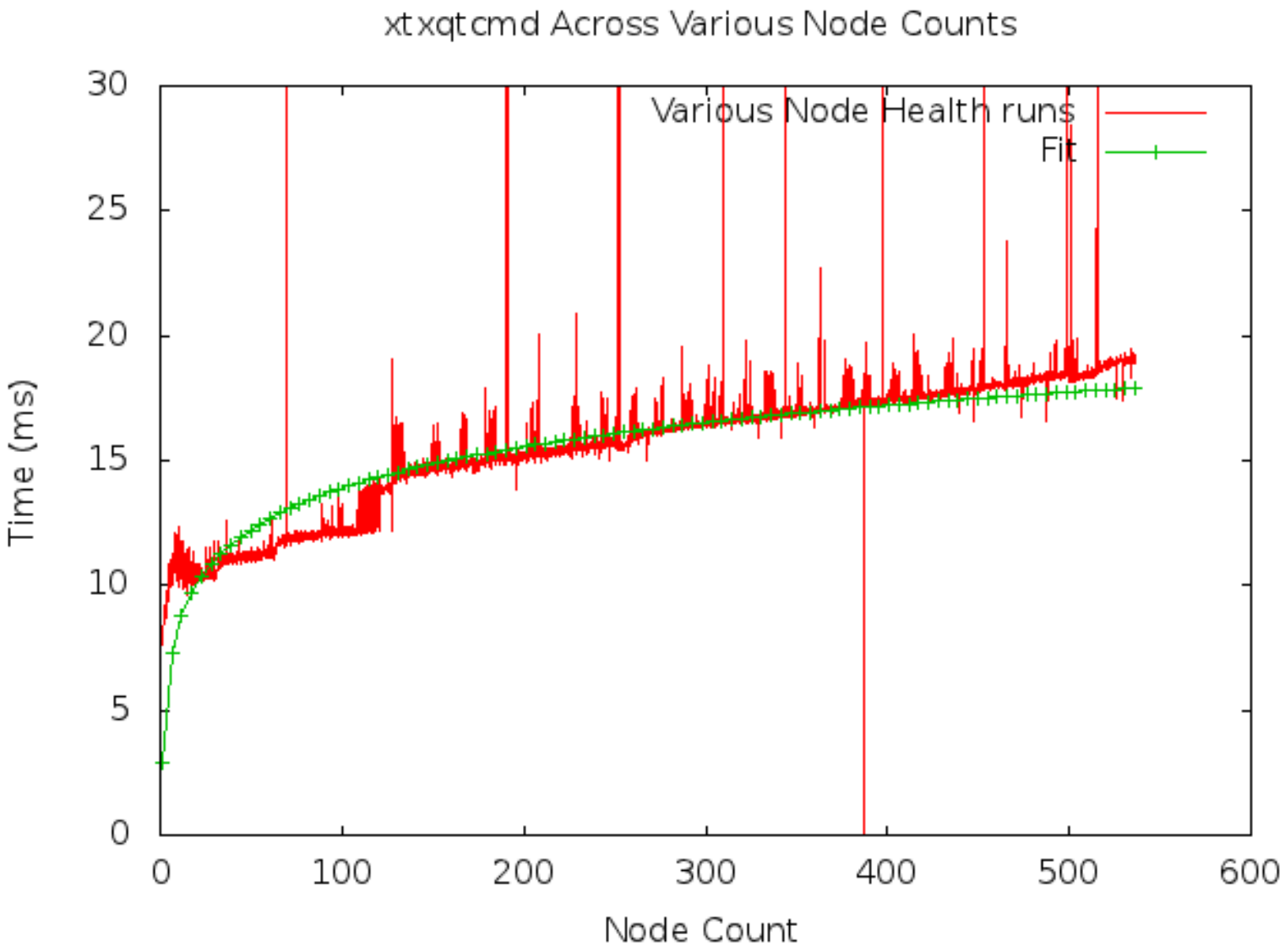
- **Gnuplot**
  - Open source graphing and curve fitting tool.
  - http://www.gnuplot.info/
  - All graphs and equations produced using Gnuplot

Using xtxqtcmd in place of xtcheckhealth

- **xtxqtcmd**
  - Uses identical fanout tree a xtcheckhealth without any overhead such as database calls or configuration file reading
  - Call xtnhc in the same manner as xtcheckhealth, without overhead



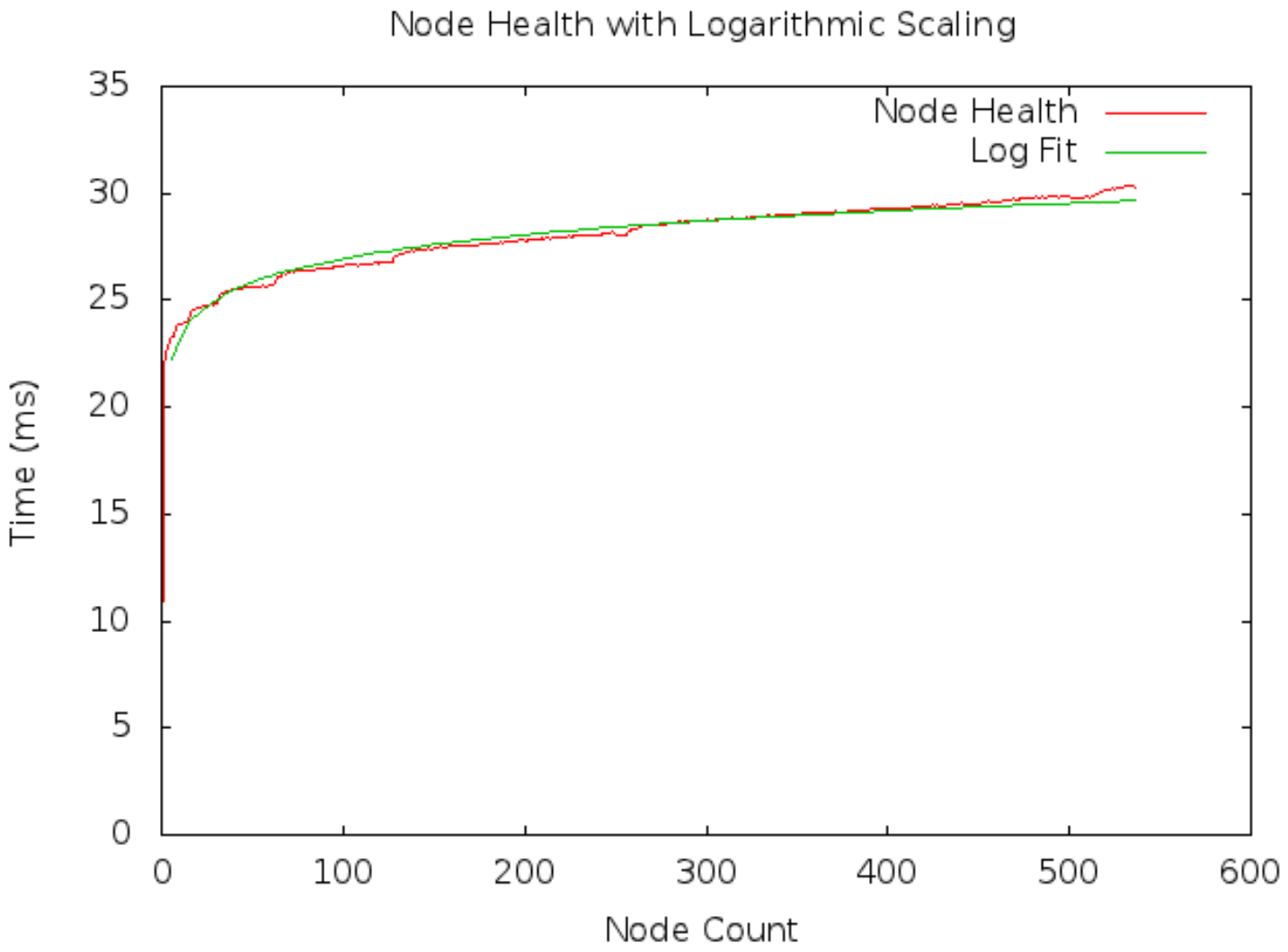
xtxqtcmd calling xtnhc in place of xtcheckhealth



xtcheckhealth as source of linearity

- **Xtxqtcmd results show xtcheckhealth caused linearity**
- **Instrumentation leads to source of linearity**
- **Call to rca-helper binary**
  - Converts nids to cnames
  - Takes ~3ms to run
  - Is called for each node
- **The fix**
  - Cache all information in one call on startup
  - Construct nid to cname relationship from data

NHC with Scaling Fix

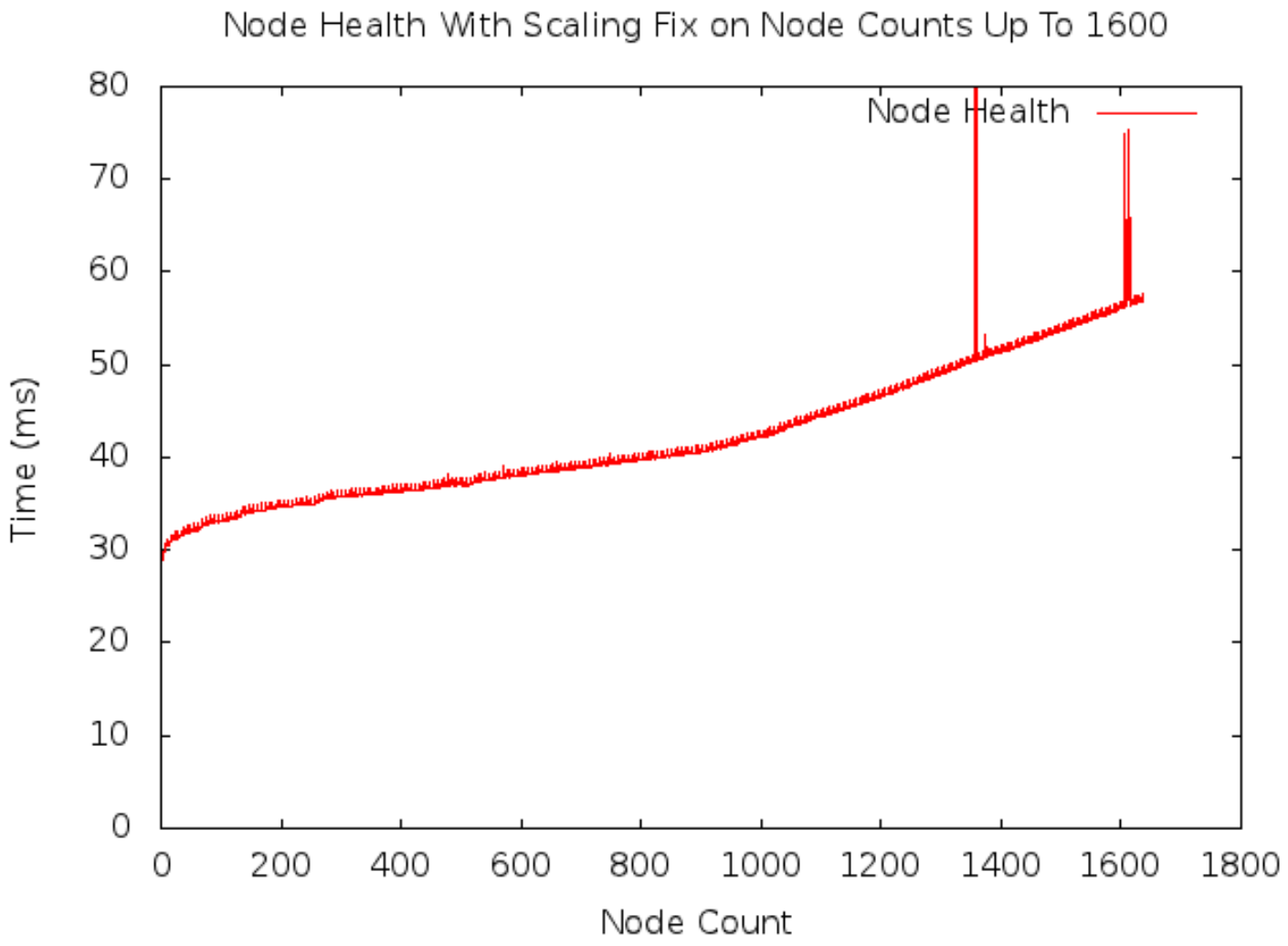


$$t = 1.61 \cdot \ln(n) + 19.6$$

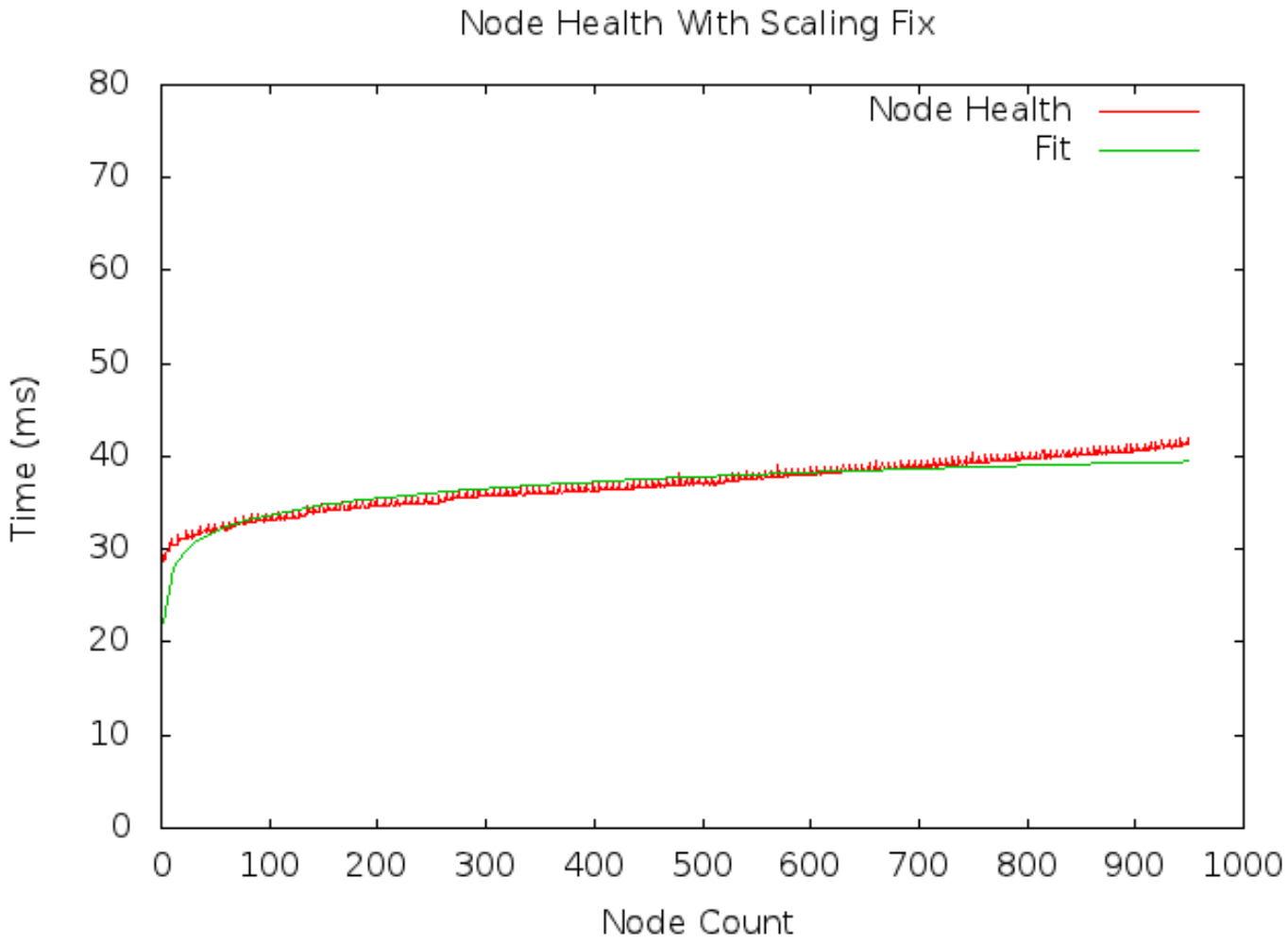
Further Testing

- Large in house system used for scaling
- System had up to 1600 nodes
- Used to verify proper scaling

NHC with Scaling Fix Up to 1600 Nodes

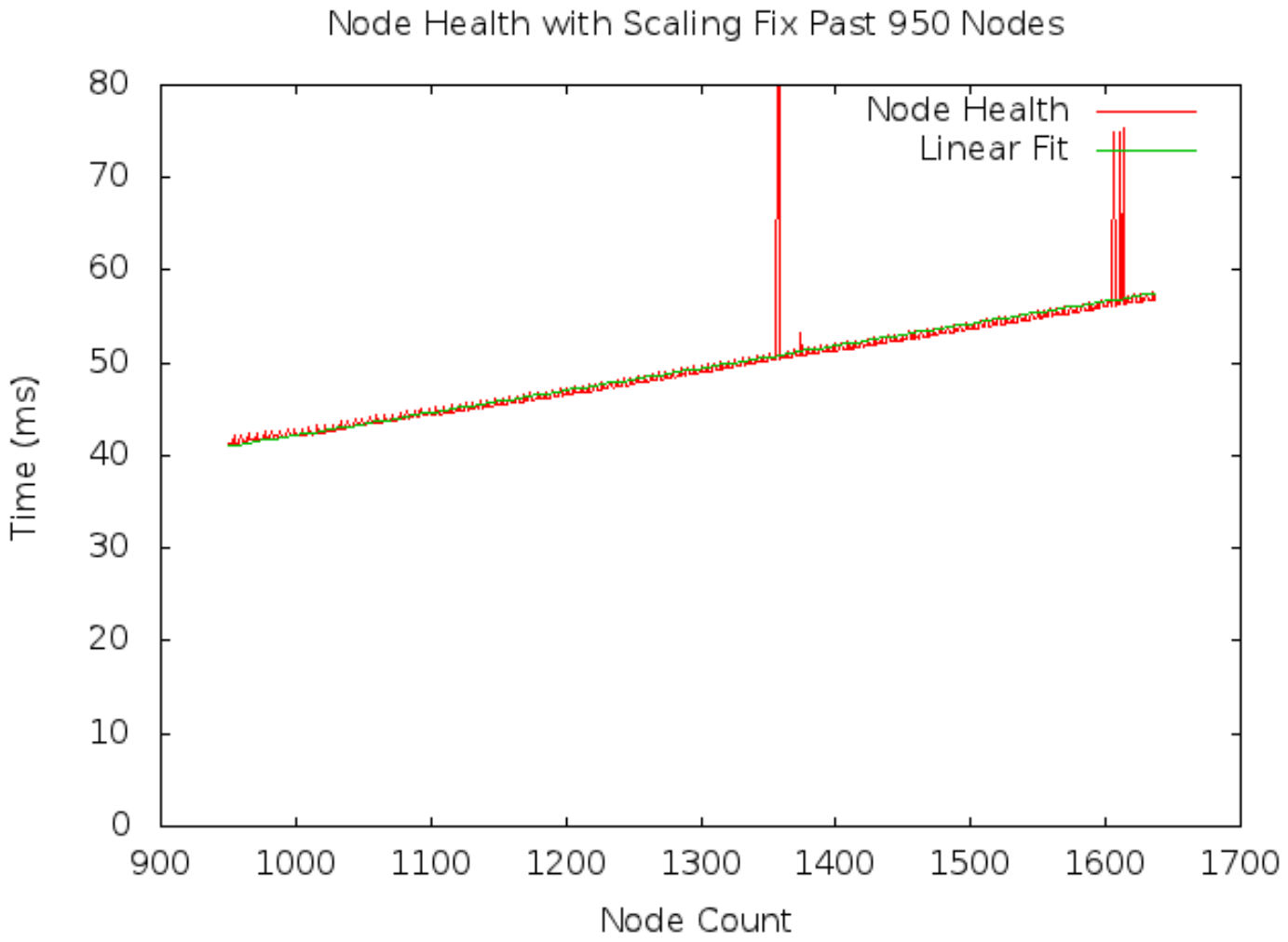


NHC on node counts up to 960



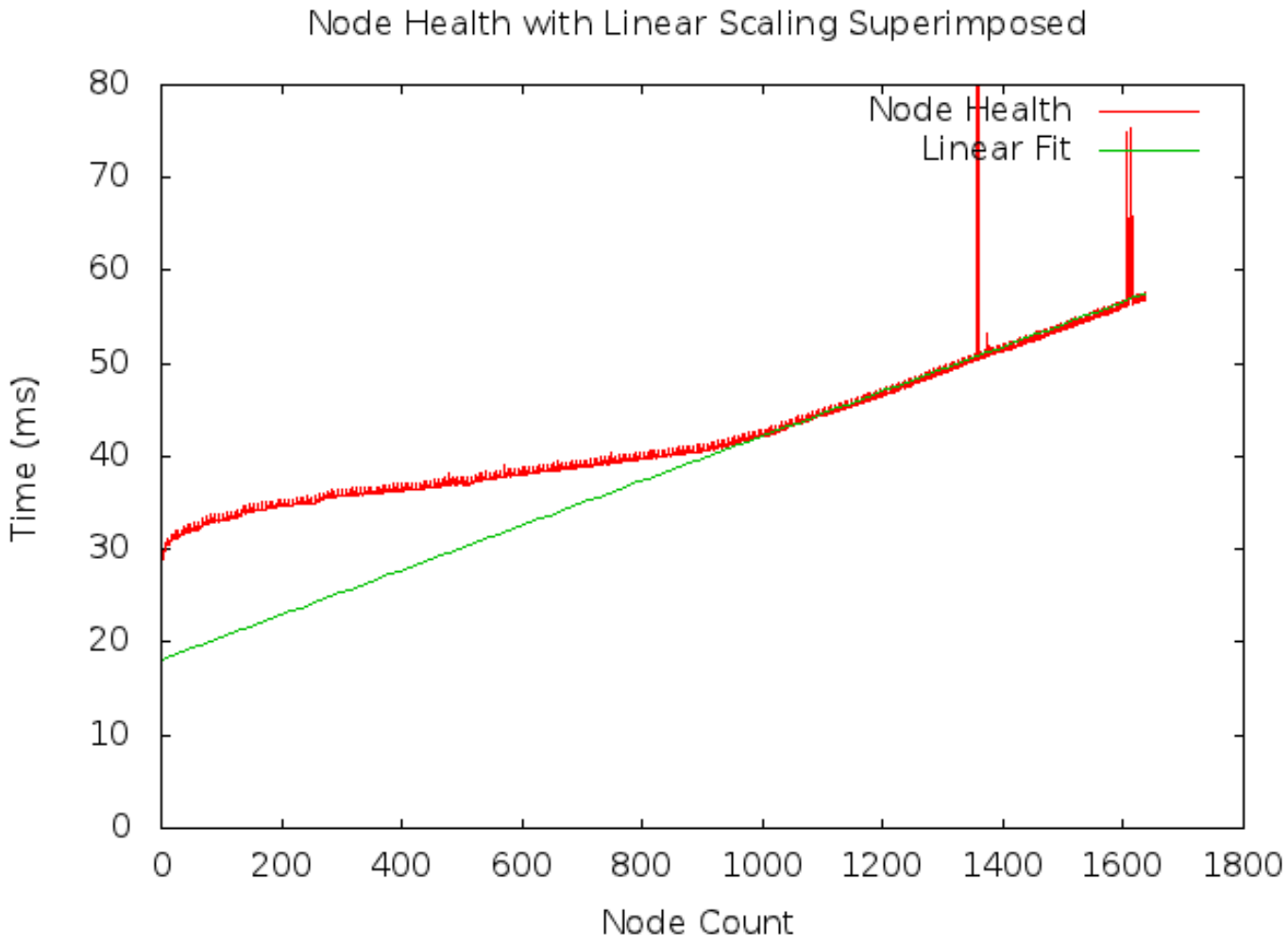
$$t = 2.54\ln(n) + 22.0$$

NHC on node counts greater than 950



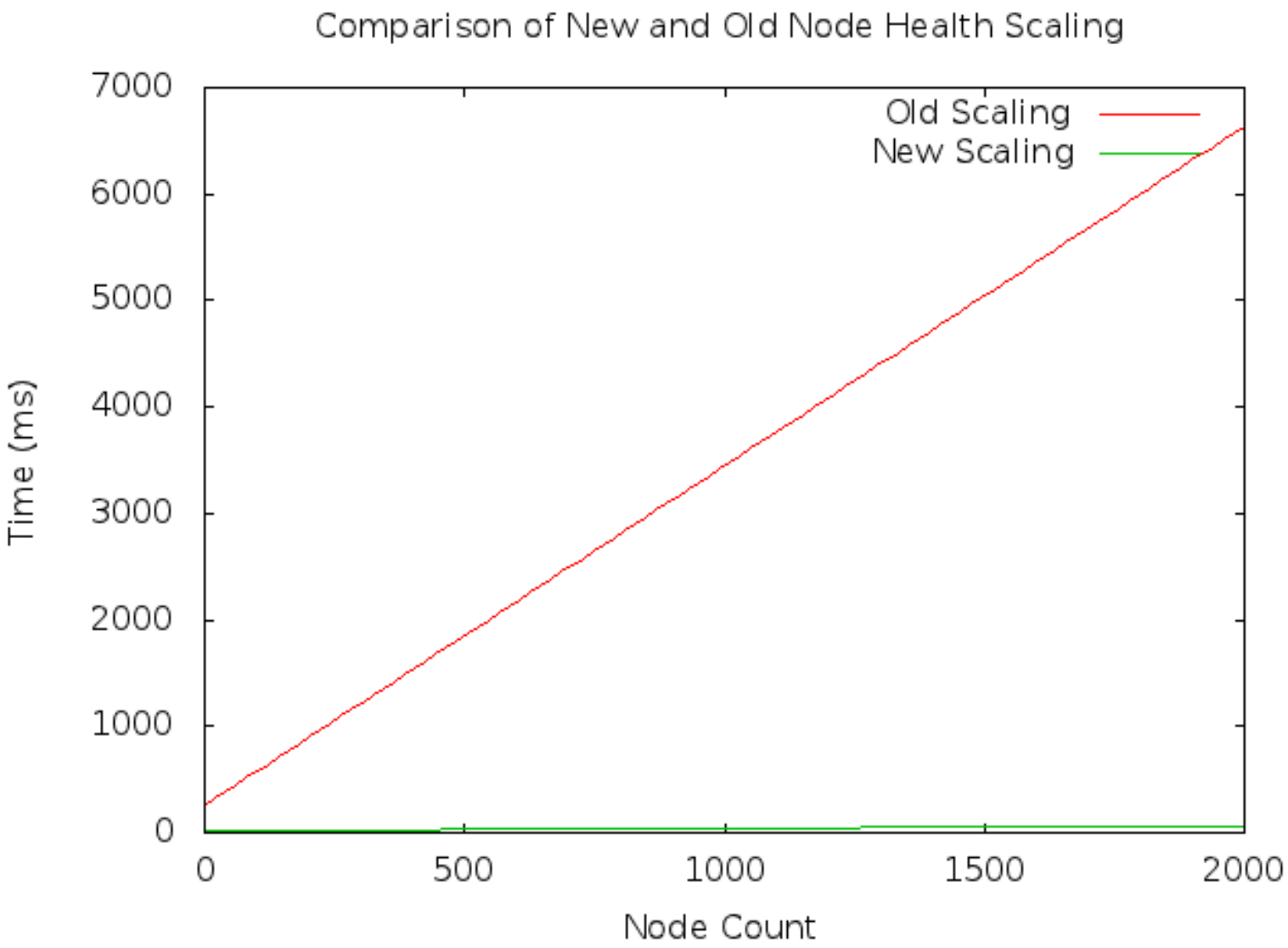
$$t = 0.024n + 18.6$$

NHC with superimposed linear fit



$$t = 0.024n + 18.6$$

Comparison of New and Old Scaling



Scaling Comparison

- New Scaling is around 100x faster

Node Count	Old Scaling (ms)	New Scaling (ms)
1000	3400	42
10000	32200	258
20000	64100	498

Automatic Dump and Reboot

- **Design Goals**
  - Reduce manual dumping and rebooting
  - Integrate into NHC
- **Dump and Reboot Info**
- **Dumpd Component Overview**
  - dumpd
  - executor
  - Configuration file
- **Use Cases**

What is a dump?

- Copy of kernel memory on a node
- Useful for debugging
- Can be used to recreate error state of node
- Non-maskable interrupt (NMI) often sent before dump taken

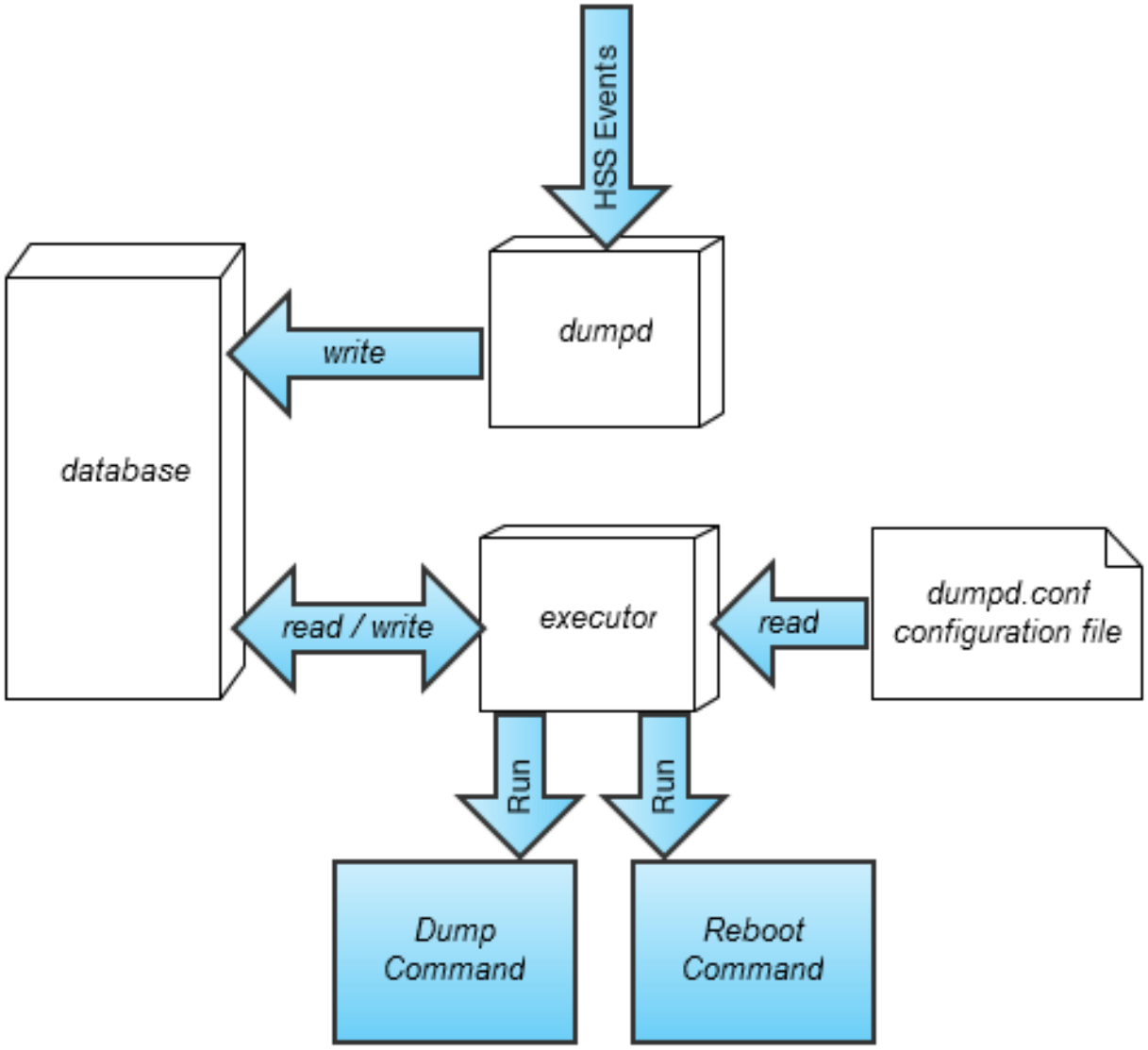
Reboot Information

- Performed by calling xtbootsys with –reboot
- Done in chunks of 50
- 50 node limit provides balance of large amount of nodes rebooted with good resiliency
  - Empirically determined

Dumpd SMW Daemon

- **NHC lives on Service / Compute Nodes**
- **Dump / Reboot commands live on SMW**
- **New SMW daemon necessary to perform commands**
- **Actions:**
  - comma-separated list of commands to perform in order
  - Examples:
  - “halt,dump”
  - “halt,dump,reboot”
  - “reboot”

SMW Dumpd Diagram



dumpd binary

- **Listen for events**
  - dump / reboot requests
  - boot events
- **Add requests to database**
- **Start executor**

database

- Stores requests
- Used in recovery
- Communication mechanism between dumpd, executor
- Stores pertinent information about request
  - cname
  - actions
  - apid
  - requester (usually service node hostname)
  - partition

executor

- Python script
- Reads in dumpd.conf config file
- Actually performs dumps and reboots

dumpd.conf

- Found at /etc/opt/cray-xt-dumpd/dumpd.conf on SMW
- Defines what is meant by ‘halt’, ‘dump’, ‘reboot’
- Halt example:

```
[halt]
command: xtnmi --partition $partition $cname
max_cnames: unlimited
timeout: 60
```

- $cname resolves to comma-separated list of cnames in actual call

NHC Integration

- **New NHC Actions**
  - Dump
  - Reboot
  - DumpReboot
- **New actions are triggered in identical manner to ‘Log’, ‘Admindown’**
- **Dumpd requests are made using HSS events**
  - Include dumpd action string
    - Ex. “halt,dump,reboot” or “reboot”
  - Include cname of node to perform action on

Using Dumpd Without NHC

- dumpd-request on SMW and service nodes
- Send dumpd requests
- Define custom action in dumpd.conf
- dumpd-request can send custom action as well

Use Case 1: Nodes Fail ‘Dump’ NHC Test

- Example would be dump on application failure
- NHC Config file entry

```
dumpdon: on
maxdumps: 3
Application: Dump 240 300 0
```

Use Case 1: Nodes Fail 'Dump' NHC Test

- ALPS calls NHC on nids 100-199
- All fail the Application test
- Nids set to "admindown"
- Nids 135, 148, and 188 are chosen at random
- Dumpd request with actions "halt,dump" are made for those nids
- Dumpd runs the "halt" action on all three nodes and then "dump" serially on each node

Use Case 2: Nodes Fail ‘Reboot’ NHC Test

- Example would be reboot on low memory situation
- NHC Config file entry

dumpdon: on
maxdumps: 3
Memory: Reboot 20 30 30 1000

Use Case 2: Nodes Fail ‘Reboot’ NHC Test

- ALPS calls NHC on nids 100-199
- All fail the Memory test
- Nids set to “unavail”
- Dumpd request with actions “reboot” are made for those nids
- Dumpd runs reboot on the first chunk of 50, then the next chunk

Use Case 3: Nodes Fail ‘DumpReboot’ NHC Test

- Example would be compute node ALPS daemon failing
- NHC Config file entry

dumpdon: on
maxdumps: 1
Alps: DumpReboot 30 60 30

Use Case 3: Nodes Fail 'DumpReboot' NHC Test

- ALPS calls NHC on nids 100-199
- All fail the Alps test
- Nids set to "unavail"
- Dumpd requests action "halt,dump,reboot" for one nid
- Dumpd request with actions "reboot" are sent for the rest
- Dumpd runs halt, then dump for the one nid
- Dumpd runs reboot on the first chunk of 50
- Dumpd runs reboot on second chunk, including dumped node
- Nodes go back to 'up' as they reboot

Use Case 4: Using Dumpd to Shut Down Nodes Prior to Reboot

- Define the following in dumpd.conf:

```
[shutdown]
Command: xtcli shutdown $cname
max_cnames: 50
simultaneous: 1
accumulation_time: 1
timeout: 60
```

- And add the following to the ‘reboot’ definition in dumpd.conf

```
pre: shutdown
```

- All ‘reboot’ actions will have a ‘shutdown’ action preceding them

Admin Reboot Collisions

- Dumpd listens to events signifying that a boot has started on a node
- That node is removed from dumpd's queue
- Multiple simultaneous dumps can be taken
- Dumpd makes no effort to avoid simultaneous dumps

Practices to Avoid

- **Specifying ‘Reboot’ action for test that will not be solved by a reboot**
  - Could lead to continuous node reboots
- **Setting maxdumps too high**
  - Dumps can use a lot of SMW space

Conclusion

- Isolation and fix of NHC scaling issue allowed normal mode run time to be decreased by up to two orders of magnitude

- Automatic dumping ad rebooting help increase automation of common admin tasks

Questions?