CRAY
THE SUPERCOMPUTER COMPANY

Gemini Software Development Using Simulation

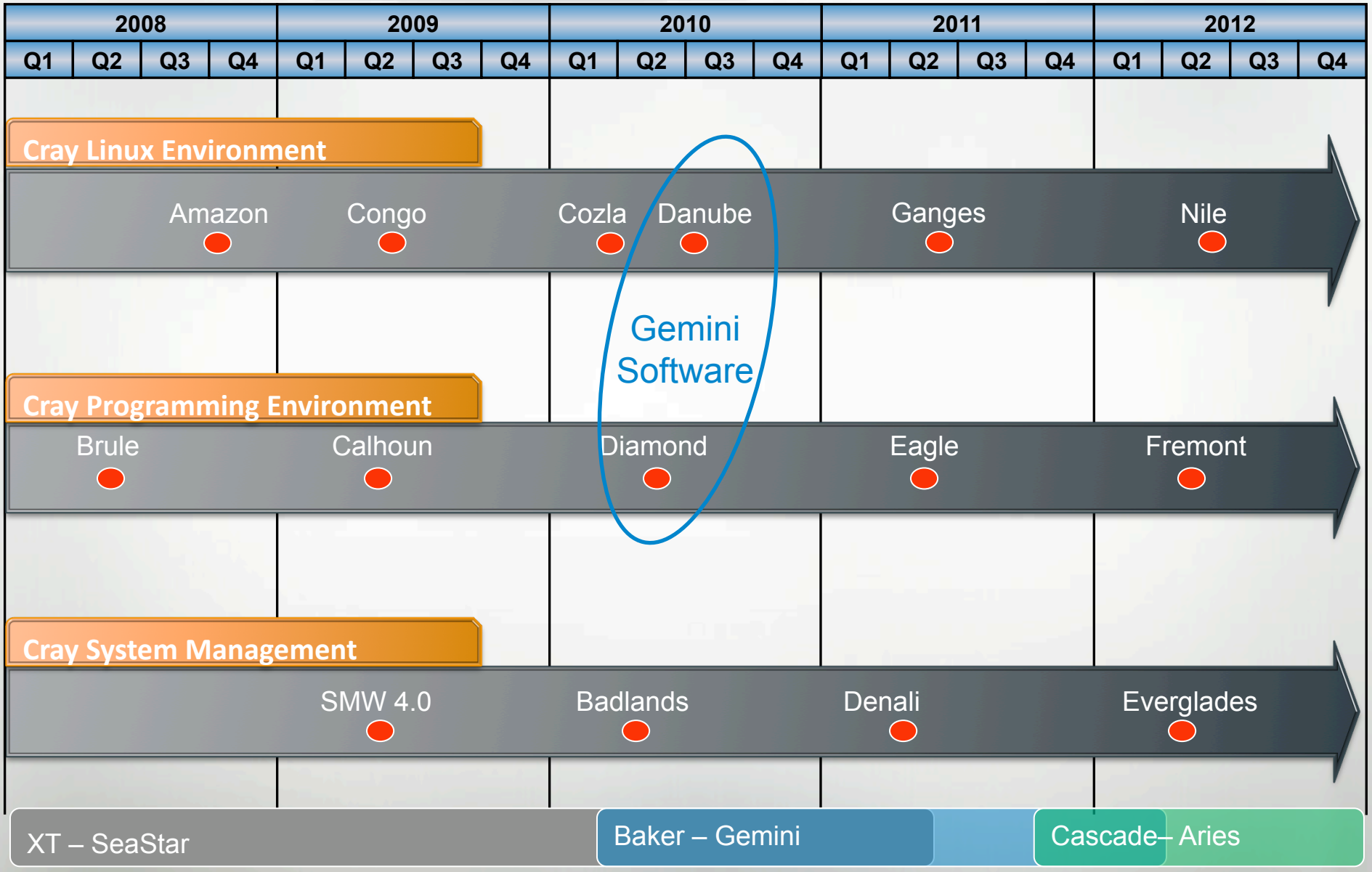
Kevin Peterson

May 7, 2009

Simulation Overview

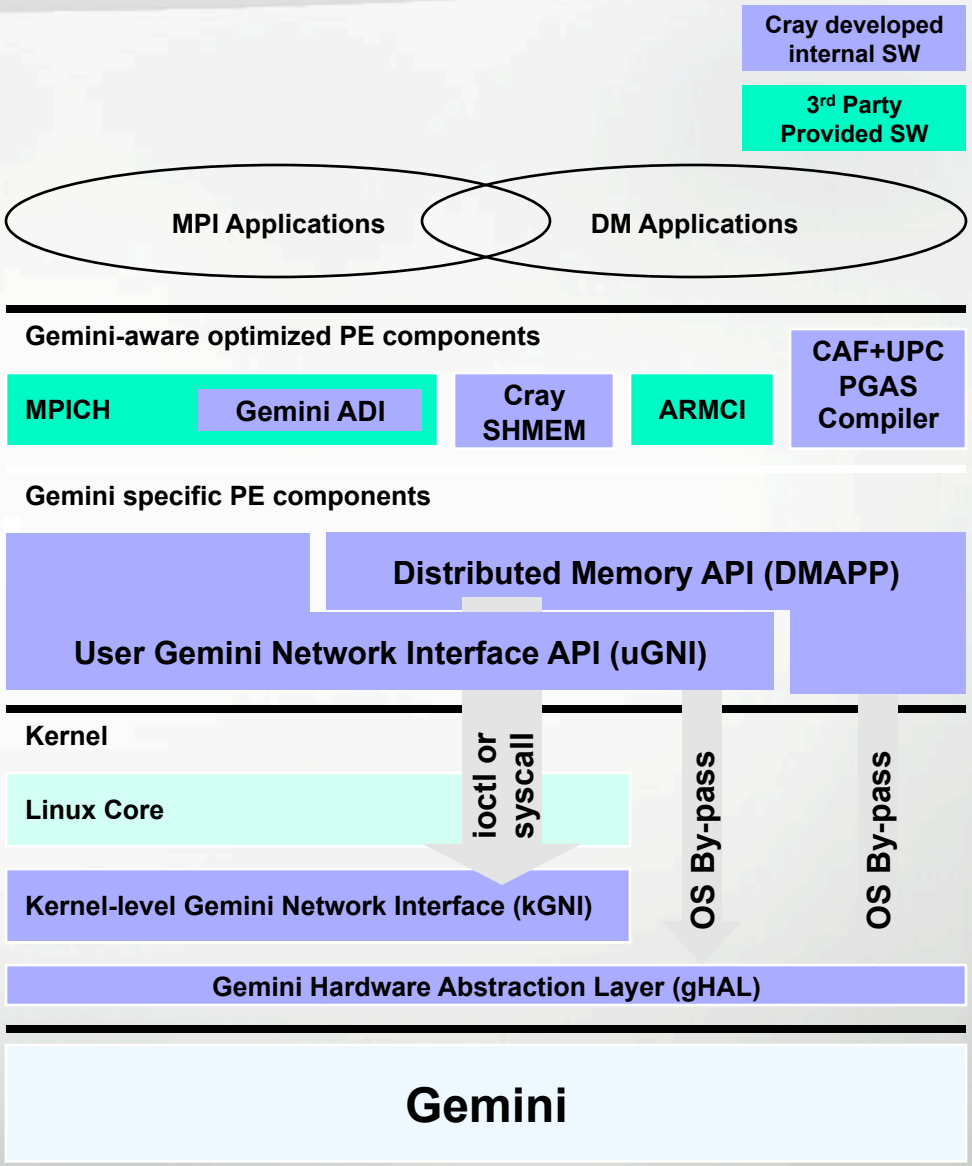
- Gemini Software
  - Cray Software Roadmap, Gemini software stack
- Simulation Goals
  - Rationale for building simulation framework
- Simulation Environment
  - SimNow ™ overview, Donut node, Lustre batch system
- Debug under Simulation
  - Debug example, Lustre routing
- Testing under Simulation
  - Sample test logs and test summaries
- Simulation Summary
  - Accomplishments, Impact on Hardware bring-up

Cray Software Roadmap



Gemini Software Stack

- Two Published APIs
  - DMAPP – shared memory
  - uGNI – message passing
- MPI Support
  - MPICH
  - Gemini Abstract Device Interface (ADI)
- SHMEM Support
  - Port to DMAPP
- PGAS Support
  - UPC
  - CAF
- Global Array Support
  - ARMCI

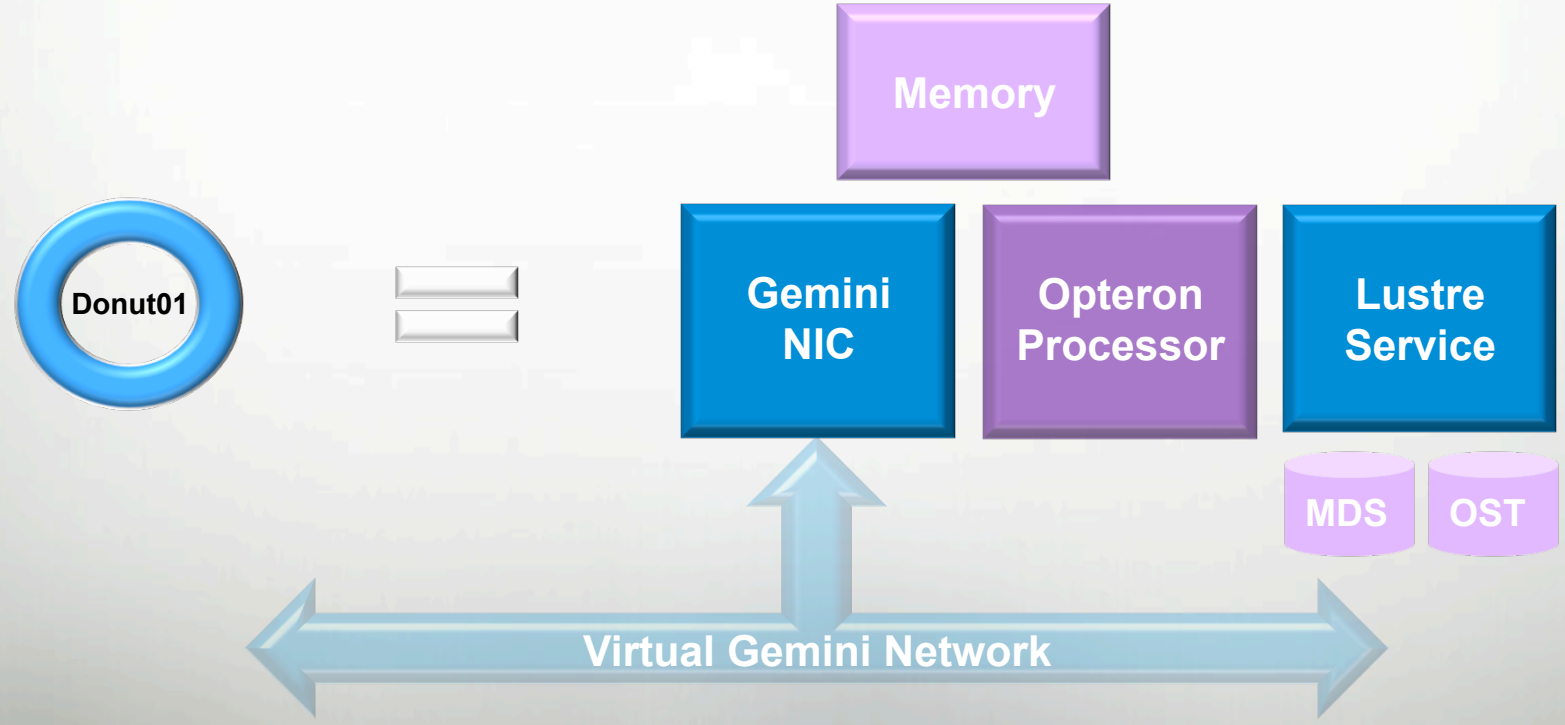


Simulation Goals & Motivation

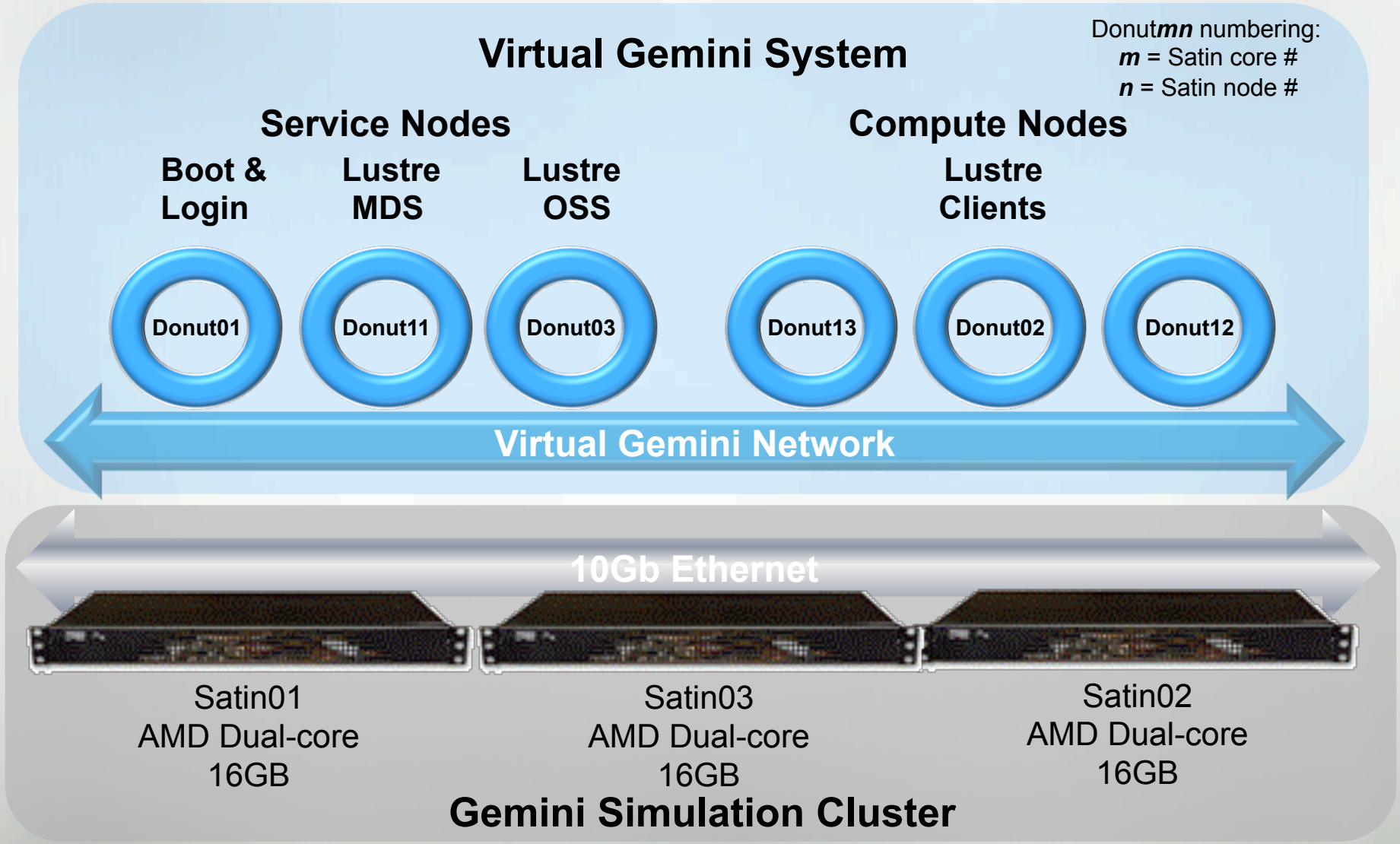
- Insure Gemini Software is running prior to HW
  - Develop Gemini HAL & Linux Driver
  - Develop Linux IP over Gemini (IPoG) & Gemini Lustre (GNI LND) drivers
  - Develop message passing & shared memory APIs
  - Develop all Gemini specific compute & server OS software (ALPS, …)
  - Test all software as a virtual XT system running "real" applications
- Reduce software debug on Gemini hardware
  - Find as many SW bugs before HW arrives
- Focus on functional correctness
  - Non-goal: performance tuning
- Lay groundwork for Aries simulation environment for Cascade
  - Gemini APIs will remain the same
  - Much of the software & infrastructure is re-usable

Gemini virtual node, a Donut!

- SimNow ™ is AMD's x86-64 functional system-level simulator
  - Provides processor, memory, and optional I/O "devices"
- A Donut is a SimNow™ instance of a Gemini node
  - Opteron processor & memory
  - Gemini NIC "device model" with virtual HSN connection
  - Optional I/O: Boot, Login, Lustre Service – Virtual disks for MDS & OSTs



Gemini Simulation Environment



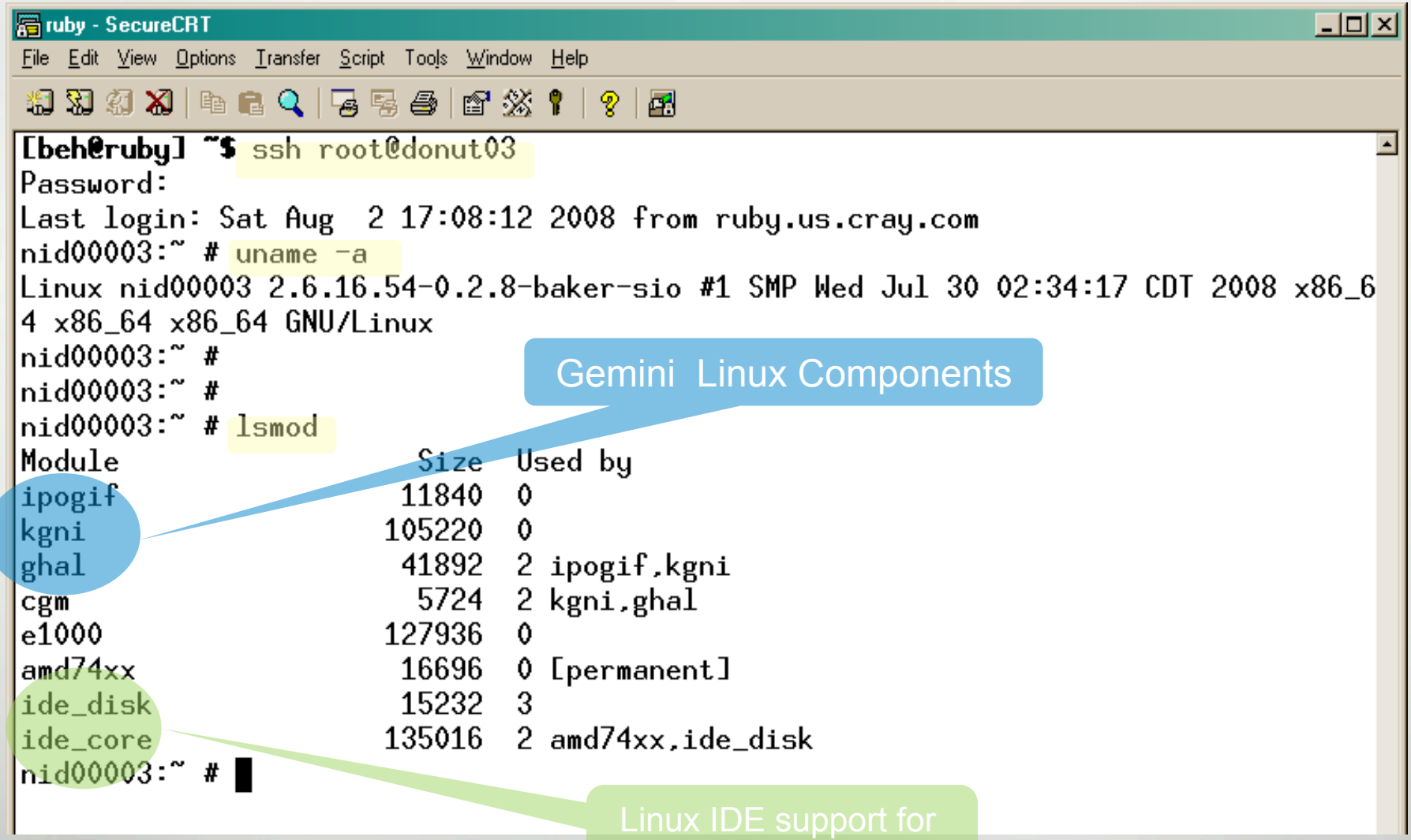
Donut Node Management Summary

Sim Host (1)	Sim Name	User	PE (2)	Server Port (3)	Sim MAC (4)	Mediator String (on sim host)	Mediator Port (on mixer)
satin01	donut01	abh	0x100	33001	fa:cd:01:XX:XX:01	mixer:20004	20000
satin01	donut11	abh	0x128	33011	fa:cd:01:XX:XX:0b	mixer:20044	20040
satin02	donut02	beh/albing	0x104	33002	fa:cd:01:XX:XX:02	mixer:20008	20004
satin02	donut12	thomson	0x12c	33012	fa:cd:01:XX:XX:0c	mixer:20048	20044
satin03	donut03	beh/albing	0x108	33003	fa:cd:01:XX:XX:03	mixer:20012	20008
satin03	donut13	beh/albing	0x130	33013	fa:cd:01:XX:XX:0d	mixer:20052	20048
satin04	donut04	igorodet	0x10c	33004	fa:cd:01:XX:XX:04	mixer:20016	20012
satin04	donut14	igorodet	0x134	33014	fa:cd:01:XX:XX:0e	mixer:20056	20052
satin05	donut05	ON ISCHED	0x110	33005	fa:cd:01:XX:XX:05	mixer:20020	20016
satin05	donut15	ON ISCHED	0x138	33015	fa:cd:01:XX:XX:0f	mixer:20060	20056
satin06	donut06	ON ISCHED	0x114	33006	fa:cd:01:XX:XX:06	mixer:20024	20020
satin06	donut16	ON ISCHED	0x13c	33016	fa:cd:01:XX:XX:10	mixer:20064	20060
velvet01	donut25	cda	0x160	33025	fa:cd:01:XX:XX:19	mixer:20100	20096
velvet01	donut26	cda	0x164	33026	fa:cd:01:XX:XX:1a	mixer:20104	20100
velvet01	donut27	bryceh	0x168	33027	fa:cd:01:XX:XX:1b	mixer:20108	20104
velvet01	donut28	Reserved (6)	0x16c	33028	fa:cd:01:XX:XX:1c	mixer:20112	20108
velvet02	donut29	howardp	0x170	33029	fa:cd:01:XX:XX:1d	mixer:20116	20112
velvet02	donut30	howardp	0x174	33030	fa:cd:01:XX:XX:1e	mixer:20120	20116
velvet02	donut31	monikatb	0x178	33031	fa:cd:01:XX:XX:1f	mixer:20124	20120
velvet02	donut32	monikatb	0x17c	33032	fa:cd:01:XX:XX:20	mixer:20128	20124
velvet03	donut33	khubert	0x180	33033	fa:cd:01:XX:XX:21	mixer:20132	20128
velvet03	donut34	khubert	0x184	33034	fa:cd:01:XX:XX:22	mixer:20136	20132
velvet03	donut35	godfrey	0x188	33035	fa:cd:01:XX:XX:23	mixer:20140	20136
velvet03	donut36	godfrey	0x18c	33036	fa:cd:01:XX:XX:24	mixer:20144	20140
velvet04	donut37	ON ISCHED	0x190	33037	fa:cd:01:XX:XX:25	mixer:20148	20144
velvet04	donut38	ON ISCHED	0x194	33038	fa:cd:01:XX:XX:26	mixer:20152	20148
velvet04	donut39	ON ISCHED	0x198	33039	fa:cd:01:XX:XX:27	mixer:20156	20152
velvet04	donut40	ON ISCHED	0x19c	33040	fa:cd:01:XX:XX:28	mixer:20160	20156

Batch

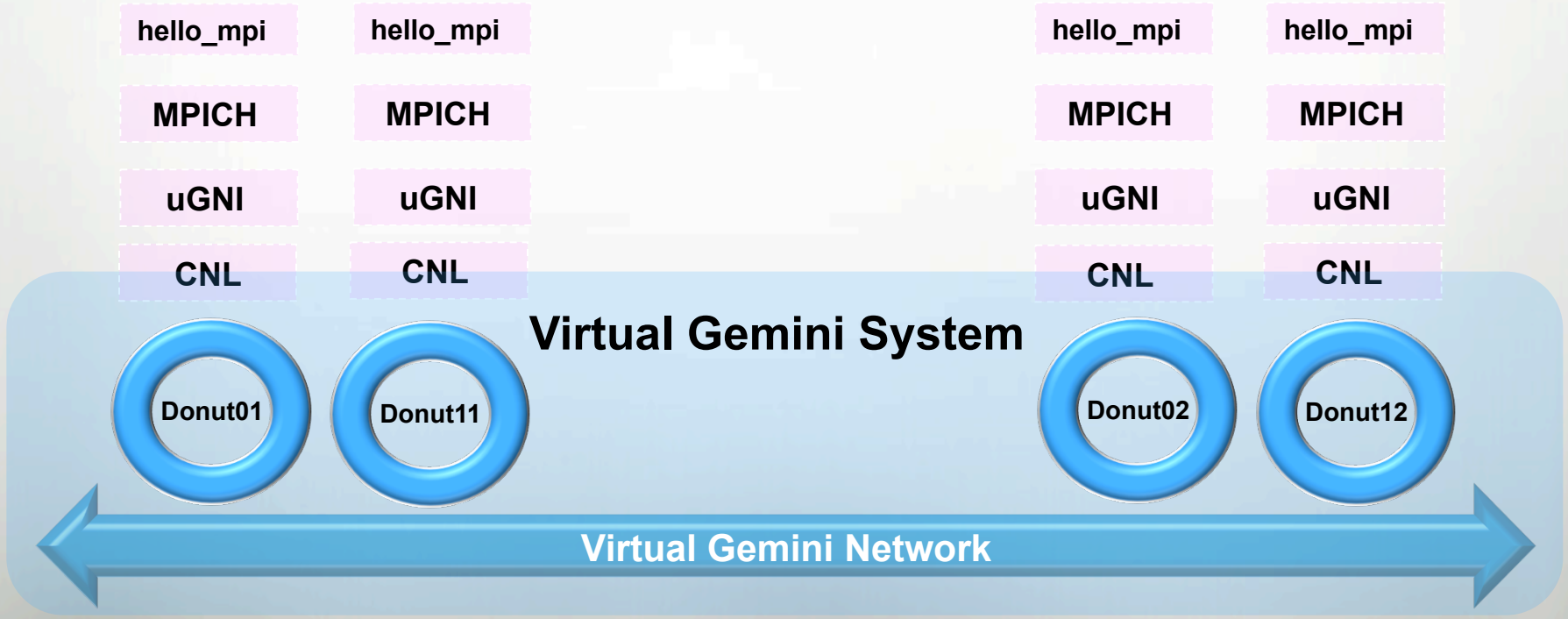
Dedicated

Sample Donut Login



Simple Simulation: MPI "Hello World" across 4 donuts

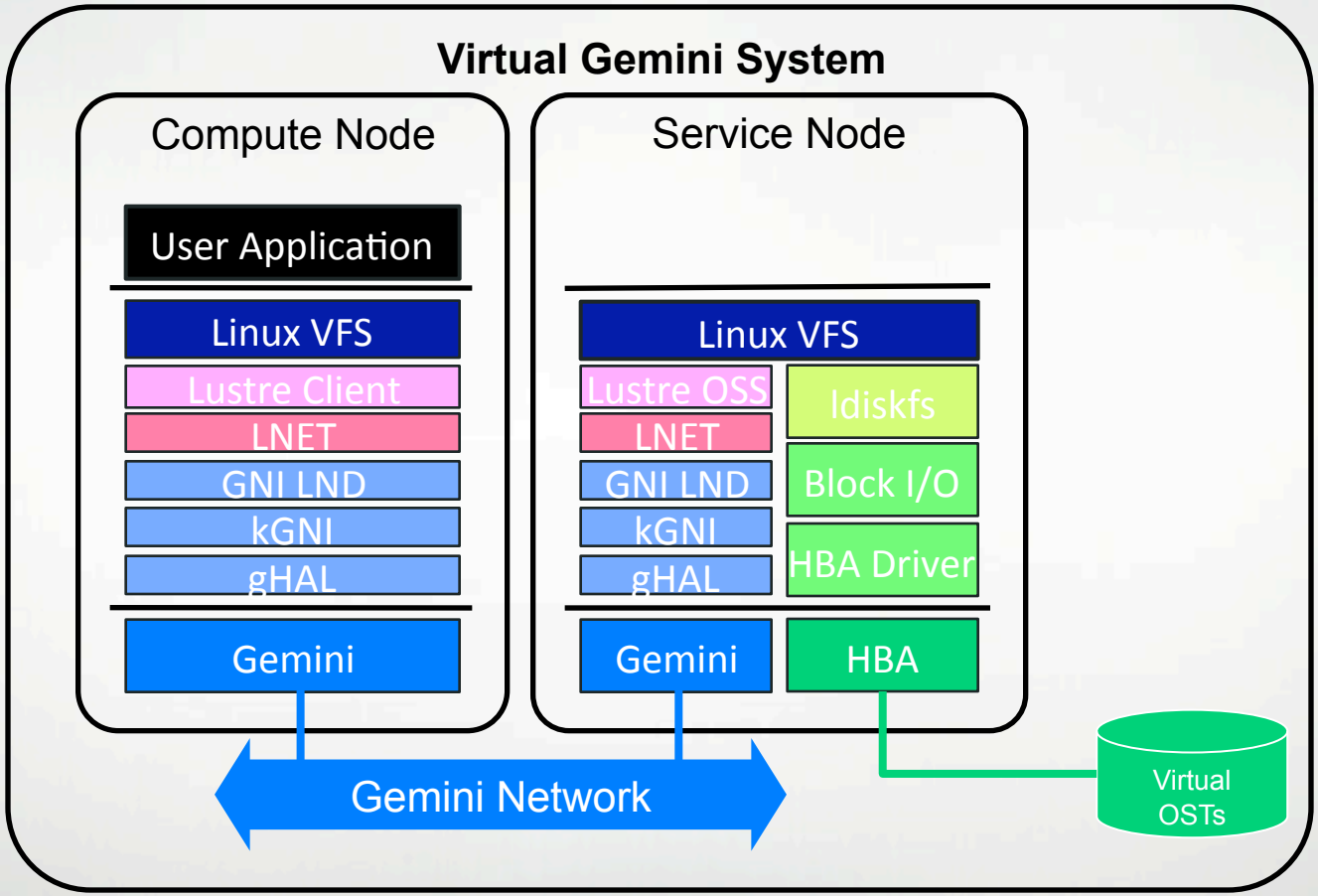
```
godfrey@nid00041:/ostest/demo> aprun -n 4 ./hello_mpi
Hello world from rank 2 of 4
Hello world from rank 1 of 4
Hello world from rank 3 of 4
Hello world from rank 0 of 4
```



Lustre Simulation Debug & Test Sessions (from Appendices)

- First example is a Lustre debug session
  - lst - Lustre self-test
    - LNET Put (write) transaction of 1MB
- Second example is a Linux test session
  - fsx - Linux file system exerciser
    - Writes large junk files to mounted Lustre scratch file system

Gemini Components Debugged in Simulation with Lustre



Example 1: Lustre Debug - LNET Self-test, *lst*

"**lst** add_test --batch write_test --concurrency 1 --loop 1 --from clients --to servers brw write check=simple size=1m"

nid00055 calls LNetPut, which generates FMA to 10.128.1.109@gni:

```
00000400:00000200:0:1235132804.358898:0:14968:0:(lib-move.c:2202:LNetPut())
LNetPut -> 12345-10.128.1.108@gni
00000800:00000200:0:1235132804.359529:0:14962:0:(gnilnd_cb.c:1494:kgnilnd_sendmsg())
$$ ffff81002d86cc00 sending FMA ffff81000159bd10 02 id 200 [ffff81001cb8a110 for 168]
msg@ffff81000159bd10 m/v/ck/pl 0be91b94/3/1b034d40/168
x30:GNILND_MSG_IMMEDIATE from 10.128.1.108@gni(1235132248969285)

00000800:00000200:0:1235132804.359602:0:14962:0:(gnilnd_cb.c:1374:kgnilnd_check_fma_send_cq())
SMSG Completed 200
```

nid00055 sees the PUT_REQ, and pushes out RDMA:

```
00000800:00000200:0:1235132804.361943:0:14962:0:(gnilnd_cb.c:1750:kgnilnd_check_fma_rx())
$$ RX on ffff81002d86cc00 from 10.128.1.109@gni
msg@ffffc200c64036d0 m/v/ck/pl 0be91b94/3/a5cf361f/0
x31:GNILND_MSG_GET_REQ from 10.128.1.109@gni(1235258178730530)

00000800:00000200:0:1235132804.361981:0:14962:0:(gnilnd_cb.c:816:kgnilnd_recv())
$$ conn ffff81002d86cc00, rxmsg ffffc200c64036d0, lntmsg ffff81001c1c2a00 niov=256
kiov=ffff81001cf76078 iov=0000000000000000 offset=0 mlen=1048576 rlen=1048576
msg@ffffc200c64036d0 m/v/ck/pl 0be91b94/3/a5cf361f/0
x31:GNILND_MSG_GET_REQ from 10.128.1.109@gni(1235258178730530)

00000800:00000200:0:1235132804.361991:0:14962:0:(gnilnd_cb.c:218:kgnilnd_setup_phys_buffer())
niov 256 offset 0 nob 1048576

00000800:00000200:0:1235132804.362030:0:14962:0:(gnilnd_cb.c:583:kgnilnd_rdma())
Post RDMA (type = 0x09) tx = 0xffff810001534400, dlvr_mode 0x0
```

nid00055 gets the GET_DONE:

```
00000800:00000200:0:1235258772.276743:0:20392:0:(gnilnd_cb.c:1750:kgnilnd_check_fma_rx())
$$ RX on ffff810035cedc00 from 10.128.1.108@gni
msg@ffffc200c6403778 m/v/ck/pl 0be91b94/3/634fa92d/0

x31:GNILND_MSG_GET_DONE from 10.128.1.108@gni(1235132248969285)
```

nid00056 gets the LNetPut, turns it into a LNetGet, and sets up PUT_REQ:

```
00000800:00000200:0:1235258772.077889:0:20392:0:(gnilnd_cb.c:1750:kgnilnd_check_fma_rx())
$$ RX on ffff810035cedc00 from 10.128.1.108@gni
msg@ffffc200c6403640 m/v/ck/pl 0be91b94/3/1b034d40/168
x30:GNILND_MSG_IMMEDIATE from 10.128.1.108@gni(1235132248969285)

00000400:00000200:0:1235258772.078301:0:20399:0:(lib-move.c:2379:LNetGet())
LNetGet -> 12345-10.128.1.108@gni

00000800:00000200:0:1235258772.078562:0:20392:0:(gnilnd_cb.c:1494:kgnilnd_sendmsg())
$$ ffff810035cedc00 sending FMA ffff810035d75910 07 id 246 [0000000000000000 for 0]
msg@ffff810035d75910 m/v/ck/pl 0be91b94/3/a5cf361f/0
x31:GNILND_MSG_GET_REQ from 10.128.1.109@gni(1235258178730530)

00000800:00000200:0:1235258772.088135:0:20392:0:(gnilnd_cb.c:1374:kgnilnd_check_fma_send_cq())
SMSG Completed 246
```

nid0056 sees the RMDA complete, and sends the GET_DONE:

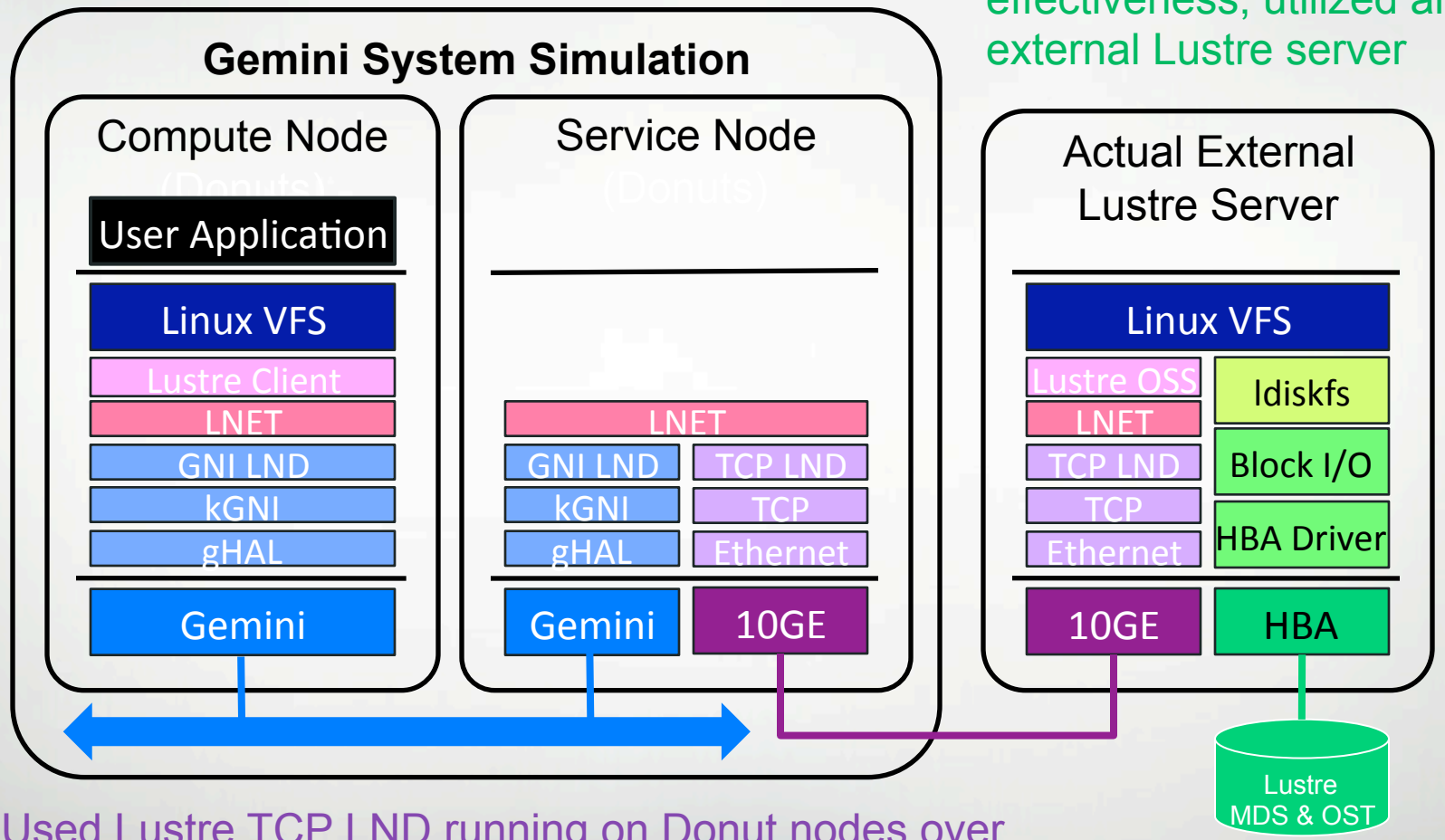
```
00000800:00000200:0:1235132804.443792:0:14962:0:(gnilnd_cb.c:1310:kgnilnd_check_rdma_cq())
RDMA completion event for tx 0xffff810001534400 type 0x09

00000800:00000200:0:1235132804.443815:0:14962:0:(gnilnd_cb.c:1494:kgnilnd_sendmsg())
$$ ffff81002d86cc00 sending FMA ffff810001534510 09 id 199 [0000000000000000 for 0]
msg@ffff810001534510 m/v/ck/pl 0be91b94/3/634fa92d/0
x31:GNILND_MSG_GET_DONE from 10.128.1.108@gni(1235132248969285)
```

nid00055 **nid00056**

Enhancing Gemini Simulation with External Lustre Routing

To improve simulation effectiveness, utilized an external Lustre server



Used Lustre TCP LND running on Donut nodes over simulation host Ethernet to route to external Lustre servers

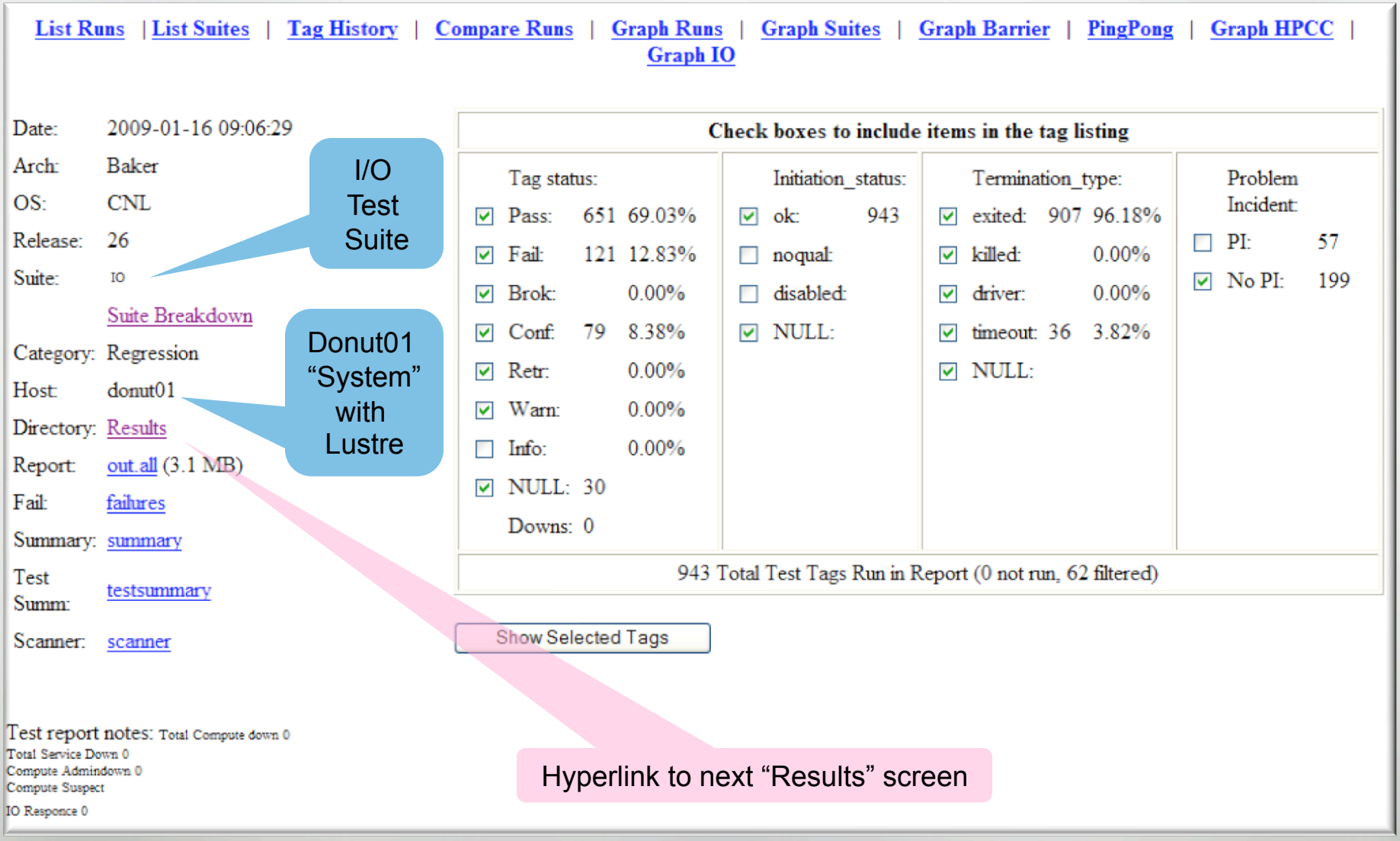
Example 2: Lustre test environment setup

```
<<<system_info_start>>>
INFORMATION ONLY: Stderr from…
/ostest/dev.gemini/baselinux/ostest/ROOT.latest/bin…
/rts_config -O OS=CNL -O ARCH=XT3:
=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=+=
Start time: 09:06:30
----------------------------------------------------------------
ENVIRONMENTAL INFORMATION:
----------------------------------------------------------------
CURRENT WORKING DIRECTORY = /lus/scratch/godfrey/tmp
RTS = /ostest/dev.gemini/baselinux/ostest/ROOT.latest
. . .
----------------------------------------------------------------
APTRUN ENVIRONMENTAL INFORMATION:
----------------------------------------------------------------
R_APP_RUN_NPES =
. . .
R_ARCH_TYPE = XT-alps
. . .
----------------------------------------------------------------
UBRUN ENVIRONMENTAL INFORMATION:
----------------------------------------------------------------
UB_CONFEQFAIL =
. . .
TESTDIR = /lus/scratch/godfrey/tmp
----------------------------------------------------------------
```

Example 2: Lustre test running *fsx* (Linux file system exerciser)

```
<<<test_start>>>
tag=CL_LTPFSX027 stime=1232118631
cmdline= "ubrun -t -x -t -D -T CL_LTPFSX027 -e LTPROOT_CL …
  aptrun -n 1 LTPROOT_CL=testcases/bin/fsx . . .
  -linux -d -l 500000 -r 4096 -t 2048 -w 2048 -W -N 10000 junkfile0.000000"
contacts="darason"
analysis=cuts
initiation_status="ok"
<<<test_output>>>
ubrun: Env LTPROOT_CL=/ostest/dev.gemini/xtcnl/ltp/ROOT.latest
ubrun: Execute Cmd: '  aptrun -n 1
/ostest/dev.gemini/xtcnl/ltp/ROOT.latest/testcases/bin/fsx
-linux -d -l 500000 -r 4096 -t 2048 -w 2048 -W -N 10000
junkfile%f'
CL_LTPFSX027    1  PASS : No failures found with the command 'aptrun'
+ The return value was 0 as expected.
<<<execution_status>>>
duration=241 termination_type=exited termination_id=0 corefile=no
cutime=164 cstime=38
<<<test_end>>>
```

Gemini I/O Regression Testing



Gemini I/O Regression Testing (more detail)

Report Detail: 2009-01-16 09:06:29
Host: donut01
Arch: Baker
Release: 26
Category: Regression
Notes:
Total Compute down 0
Total Service Down 0
Compute Admindown 0
Compute Suspect
IO Responce 0

Suite	Test Tags	Pass	Fail	Brok	Other	Pass %	Fail %	Brok %	Other %	Filtered
REPORT TOTAL	944	651	112	0		68.96	11.86	0.00	0.00	62
IO	915	639	116		160	69.84	12.68	0.00	17.49	59

Suite: IO Status: FAIL Show tags

172 tags matched your selection

	tag	tcid	testcase	status	initiation	termination	PI	contact	Suites
6	CL_LTPFSX032	CL_LTPFSX032	1	FAIL	ok	exited		darason	IO
7	CL_LTPFSX033	CL_LTPFSX033	1	FAIL	ok	exited	TestIssue	darason	IO
8	CL_LTPFSX034	CL_LTPFSX034	1	FAIL	ok	exited	_INVESTIGATE_	darason	IO
10	CL_LTPFSX020	CL_LTPFSX020	1	FAIL	ok	exited	_INVESTIGATE_	darason	IO
11	CL_LTPFSX021	CL_LTPFSX021	1	FAIL	ok	exited	_INVESTIGATE_	darason	IO

GROMACS application running from in Lustre scratch

```
godfrey@nid00001:/lus/scratch/godfrey/RUN/d.dppc> aprun -n 4 ../mdrun
[PE_0]: inet_ipaddr_from_dev: ioctl SIOCGIFADDR call failed 19
NNODES=4, MYRANK=0, HOSTNAME=nid00004
NNODES=4, MYRANK=2, HOSTNAME=nid00006
NNODES=4, MYRANK=3, HOSTNAME=nid00007
NNODES=4, MYRANK=1, HOSTNAME=nid00005
NODEID=0 argc=1
NODEID=3 argc=1
NODEID=2 argc=1
NODEID=1 argc=1
                         :-)  G  R  O  M  A  C  S  (-:

               GRoups of Organic Molecules in ACtion for Science

                            :-)  VERSION 3.2.1  (-:

  M E G A - F L O P S   A C C O U N T I N G

       Based on real time for parallel computer.
   RF=Reaction-field  Free=Free Energy  SC=Softcore
   T=Tabulated        S=Solvent         W=Water     WW=Water-Water

          Computing:      M-Number      M-Flop's  % Flop's
                 LJ    280.116462   8683.610322     9.9
            Coulomb    241.217644   6512.876388     7.5
                                        .
                                        .
                                        .
```

GROMACS application running from in Lustre scratch (continued)

```
                                .
                                .
                                .
             Propers      1.758208    402.629632     0.5
          Impropers       0.310272     64.536576     0.1
        RB-Dihedrals      2.482176    613.097472     0.7
              Virial     12.318364    221.730552     0.3
              Update     12.307456    381.531136     0.4
             Stop-CM     12.185600    121.856000     0.1
           Calc-Ekin     12.429312    335.591424     0.4
               Lincs      5.168128    310.087680     0.4
           Lincs-Mat     72.142848    288.571392     0.3
             Shake-V     12.307456    184.611840     0.2
           Shake-Vir     12.307456    221.534208     0.3
              Settle      2.425856    783.551488     0.9
         Total                  87334.11690   100.0

               NODE (s)   Real (s)      (%)
       Time:    128.000    128.000    100.0
                      2:08
               (Mnbf/s)   (MFlops) (ps/NODE hour) (NODE hour/ns)
Performance:     18.772    682.298      5.625    177.778

gcq#132: "It's Not Your Fault" (Pulp Fiction)

Application 251 resources: utime 0, stime 0
godfrey@nid00001:/lus/scratch/godfrey/RUN/d.dppc>
```

General Regression Testing on Donut01

List Runs | List Suites | Tag History | Compare Runs | Graph Runs | Graph Suites | Graph Barrier | PingPong | Graph HPCC | Graph IO

Date: 2008-11-24 09:01:24
Arch: BAKER
OS: CNL
Release: 26
Suite: APP COARRAY CUST INTRCNCT IO MPI NET OS SCHED SHMEM UPC
Suite Breakdown
Category: Regression
Host: donut01
Directory: Results
Report: out.all (19.7 MB)
Fail: failures
Summary: summary
Test Summ: testsummary
Scanner: scanner

Test report notes: Total Compute down 11
Totla Service Down 0
Compute Admindown 0
Compute Suspect 0
IO Responce 0

Check boxes to include items in the tag listing

Tag status:			Initiation_status:		Termination_type:			Problem Incident:	
☑ Pass:	294	23.37%	☑ ok:	1258	☑ exited:	1221	97.06%	☐ PI:	85
☑ Fail:	61	4.85%	☐ noqual:		☑ killed:		0.00%	☑ No PI:	842
☑ Brok:		0.00%	☐ disabled:		☑ driver:		0.00%		
☐ Conf:	791	62.88%	☑ NULL:		☑ timeout:	37	2.94%		
☑ Retr:		0.00%			☑ NULL:				
☑ Warn:		0.00%							
☐ Info:		0.00%							
☑ NULL:	25								
Downs:	0								

1258 Total Test Tags Run in Report (0 not run, 87 filtered)

Show Selected Tags

Comprehensive Test Suite

APP	COARRAY
CUST	INTRCNCT
IO	MPI
NET	OS
SCHED	SHMEM
UPC	

Gemini Simulation Summary

- Top to bottom Gemini software stack running across multiple Gemini nodes
  - MPI, UPC, & CAF applications
  - Tests, diagnostics & I/O exercisers
  - uGNI, DMAPP
  - IP over Gemini (IPoG)
  - Gemini Network Interface (GNI) LND
  - Kernel Gemini Network Interface (kGNI)
  - Gemini Hardware Abstraction Layer (gHAL)
- Validates Gemini software components
  - Numerous bugs identified and resolved
  - System integration issues worked out prior to hardware
- Minimizes hardware system integration
  - Compute nodes, service nodes, I/O all tested as "system"
- Code executing on Gemini prototype system with few changes!

Acknowledgements & References

- Simulation Infrastructure
  - Bryan Hardy and Cray IT support team
- Test Infrastructure, screen shots, & log files
  - Jason Godfrey and Cray test team
- Gemini Software
  - Cray Gemini software team
- Gemini Lustre I/O Configurations
  - John Carrier
- Cray Gemini Test environment
  - OSTEST Page: http://insidecray.mw.cray.com/~tests/
  - RTS Page: http://insidecray.mw.cray.com/~tests/rts/
- AMD x86-64 SimNow™
  - Website: http://developer.amd.com/cpu/simnow/Pages/default.aspx

Thank you!

kpeterso@cray.com


CRAY
THE SUPERCOMPUTER COMPANY