

Rx Listener Performance

or: How to Saturate a 10GbE Link with an OpenAFS Rx File-server

Andrew Deason

June 2019

OpenAFS Workshop 2019

Overview

- Problem and background
- Baseline: ~1.7 gbps
- `foreach(why_are_we_so_slow):`
 - Discuss issue
 - Show solution
 - Performance impact
- End result: 10gbps+
- Other considerations, future

The Problem

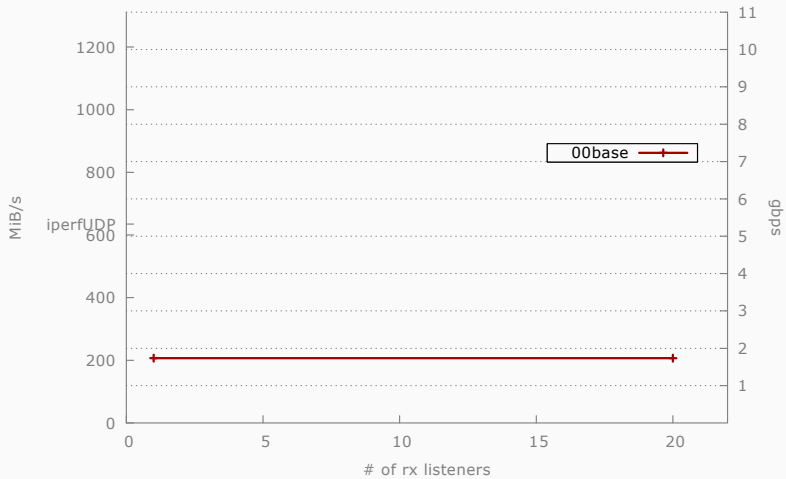
- Customer has 1G volume, files are 1M+
- Hundreds of clients, all fetch at once
- Fileserver saturated at 1-2gbps
- 1GiB * 100clients @ 1.5gbps \approx 9.5 minutes
- 1GiB * 100clients @ 10gbps \approx 1.5 minutes

- We do **not** care about:
 - Single-client performance, latency
 - Uncached files
 - Complex solutions (DPF, TCP)
 - Other servers

Test Environment

- Fileserver
 - Solaris 11.4
 - HP ProLiant DL360 Gen9
 - Xeon E5-2667v3, 8/16 cores
- Clients
 - Fake afscp clients on Debian 9.5
 - HP ProLiant DL360 Gen10
 - Xeon Gold 6136, 12/24 cores
- 2x Broadcom Limited NetXtreme II BCM57810 10gbps NIC
- Harness: `afscp_bench` Python script

Step 0: Baseline (master fc7e1700)



Step 0: Baseline (master fc7e1700)

```
$ prstat -c -n 20 5 6
```

PID	USERNAME	SIZE	RSS	STATE	PRI	NICE	TIME	CPU	PROCESS/NLWP
12201	root	32804K	26628K	cpu14	59	-5	0:00:46	8.298%	daifilesver/140
833	root	153620K	35596K	sleep	59	0	3:00:02	0.278%	cmd/44
925	root	14740K	5336K	sleep	59	0	0:15:16	0.107%	syslogd/10
5	root	0K	0K	sleep	99	-20	1:37:12	0.027%	zpool-rpool/191
787	root	221412K	50420K	sleep	59	0	2:23:04	0.018%	sstored/21
12192	root	26396K	19640K	sleep	59	0	0:00:00	0.015%	python3.5/1
12417	root	14392K	8596K	sleep	59	0	0:00:00	0.009%	prstat/1
12425	root	19600K	12968K	sleep	59	0	0:00:00	0.006%	perl/1
12238	root	19532K	11420K	sleep	59	0	0:00:00	0.003%	ssh/1
12225	root	19532K	11416K	sleep	59	0	0:00:00	0.003%	ssh/1
12326	root	19532K	11420K	sleep	59	0	0:00:00	0.003%	ssh/1
12223	root	19532K	11416K	sleep	59	0	0:00:00	0.003%	ssh/1
12316	root	19532K	11420K	sleep	59	0	0:00:00	0.003%	ssh/1
12235	root	19528K	11416K	sleep	59	0	0:00:00	0.003%	ssh/1
12374	root	19532K	11420K	sleep	59	0	0:00:00	0.003%	ssh/1
12209	root	13376K	6224K	sleep	59	0	0:00:00	0.003%	sh/1
12276	root	19532K	11420K	sleep	59	0	0:00:00	0.003%	ssh/1
12398	root	19532K	11420K	sleep	59	0	0:00:00	0.003%	ssh/1
12322	root	19532K	11420K	sleep	59	0	0:00:00	0.003%	ssh/1
12362	root	19532K	11420K	sleep	59	0	0:00:00	0.003%	ssh/1

Total: 248 processes, 1270 lwps, load averages: 1.41, 3.36, 4.01

“Is this server even busy?”

Step 0: Baseline (master fc7e1700)

```
$ mpstat 5 6
CPU minf mjf xcal intr ithr csw icsw migr smtx srw syscl usr sys st idl
0 0 0 39 1561 102 3146 3 179 2227 0 5462 4 8 0 88
1 0 0 1 620 0 1310 1 33 1123 0 2894 2 4 0 94
2 72 0 664 1036 0 2373 2 135 3930 0 29648 8 15 0 77
3 0 0 30 752 50 1492 1 47 1222 0 3124 2 4 0 93
4 0 0 6 1101 0 3307 2 178 2123 0 5331 4 7 0 89
5 0 0 3382 48594 48582 254 20 10 8919 0 222 0 23 0 77
6 0 0 1786 280 0 845 3 48 7803 0 75300 14 24 0 62
7 0 0 18 1244 2 5119 4 223 2865 0 7389 6 11 0 82
8 0 0 7 685 0 1434 1 41 1263 0 3314 2 5 0 93
9 0 0 9 1191 3 3083 2 172 2140 0 5024 4 7 0 89
10 0 0 449 485 0 1039 1 28 2902 0 21493 5 9 0 86
11 0 0 13 1268 0 3189 2 171 2199 0 5669 4 8 0 88
12 0 0 12 786 0 1628 2 43 1255 0 2973 2 4 0 94
13 0 0 19 1298 0 4179 3 183 2490 0 6448 5 10 0 85
14 0 0 2060 145 1 319 2 9 9479 0 87550 15 26 0 59
15 0 0 3244 48528 48518 336 27 12 9018 0 321 0 23 0 76
```

“Is this server even busy?”

Step 0: Baseline (master fc7e1700)

```
$ prstat -lm -c -n 20 5 6
PID USERNAME    USR  SYS  TRP  TFL  DFL  LCK  SLP  LAT  VCX  ICX  SCL  SIG  LWRD  PROCESS/OWNERNAME
12201 root          36,87 63,04 0,002 0,000 0,000 0,022 0,046 0,024 318  29 1037K  0    3  dafileservr/rx_listener
12201 root          0,445 0,676 0,000 0,000 0,000 98,79 0,001 0,086 355  2  3213  0    30  dafileservr
12201 root          0,444 0,670 0,000 0,000 0,000 98,80 0,000 0,081 351  2  3041  0   113  dafileservr
12201 root          0,436 0,677 0,000 0,000 0,000 98,79 0,003 0,092 332  4  2934  0   112  dafileservr
12201 root          0,441 0,666 0,000 0,000 0,000 98,82 0,001 0,069 334  6  3055  0    98  dafileservr
12201 root          0,440 0,665 0,000 0,000 0,000 98,82 0,001 0,074 341  4  2941  0   135  dafileservr
12201 root          0,436 0,663 0,000 0,000 0,000 98,83 0,001 0,067 337  0  3009  0    69  dafileservr
12201 root          0,436 0,658 0,000 0,000 0,000 98,83 0,000 0,074 342  3  3018  0    88  dafileservr
12201 root          0,431 0,657 0,000 0,000 0,000 98,83 0,001 0,081 330  4  2906  0    71  dafileservr
12201 root          0,428 0,659 0,000 0,000 0,000 98,83 0,000 0,082 334  1  3084  0    77  dafileservr
12201 root          0,427 0,659 0,000 0,000 0,000 98,82 0,002 0,090 331  6  2803  0   126  dafileservr
12201 root          0,434 0,651 0,000 0,000 0,000 98,84 0,001 0,074 349  2  2967  0    76  dafileservr
12201 root          0,439 0,644 0,000 0,000 0,000 98,84 0,000 0,073 353  3  2897  0   136  dafileservr
12201 root          0,434 0,647 0,000 0,000 0,000 98,84 0,001 0,077 355  2  2867  0    55  dafileservr
12201 root          0,429 0,649 0,000 0,000 0,000 98,85 0,001 0,072 336  1  2909  0   106  dafileservr
12201 root          0,439 0,637 0,000 0,000 0,000 98,85 0,001 0,074 355  5  2792  0    41  dafileservr
12201 root          0,427 0,647 0,001 0,000 0,000 98,84 0,001 0,084 333  5  2954  0    73  dafileservr
12201 root          0,431 0,644 0,000 0,000 0,000 98,85 0,000 0,073 331  3  2923  0    58  dafileservr
12201 root          0,426 0,643 0,000 0,000 0,000 98,83 0,000 0,099 345  2  2895  0   128  dafileservr
12201 root          0,424 0,642 0,000 0,000 0,000 98,84 0,001 0,089 383  2  2865  0    60  dafileservr
Total: 243 processes, 1265 lups, load averages: 1.50, 3.44, 4.00
█
```

One thread is doing all the work!

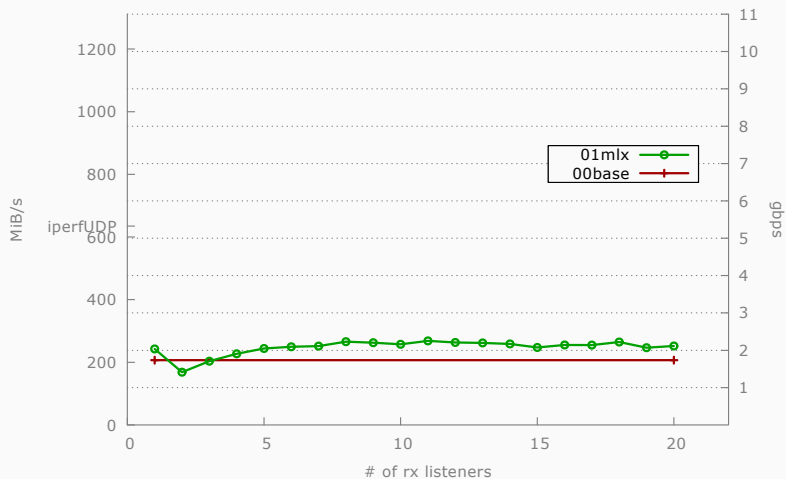
- aka `rxi_ListenerProc`, “the listener”, etc.
- TCP: `read(fd)/recv(fd)` per stream
- UDP: `recvmsg(fd)` for everyone

- Listener calls `recvmsg()`, parses, hands out data
- Other processing, too (later)
- ... for **all 128/256/etc threads** (-p)
- We're sending data, but receive ACKs

Step 1: Multiple Listeners

- Create multiple threads to run `rx_listener_proc()`
- `recvmsg()` itself internally serialized
- Everything after `recvmsg()` runs in parallel (per-conn)
 - `conn_recv_lock`
- How many threads?

Step 1: Multiple Listeners



Step 1: Multiple Listeners

```
$ mpstat 5 6
```

CPU	minf	mjf	xcal	intr	ithr	csw	icsw	migr	smtx	srw	syscl	usr	sys	st	idl
0	0	0	142	459	102	13132	148	1411	22646	0	32130	27	47	0	27
1	0	0	55	158	0	10812	135	951	17565	0	23420	19	35	0	45
2	0	0	131	24220	24079	7835	253	814	14901	0	15356	14	40	0	46
3	0	0	59	237	77	10731	140	972	17784	0	23585	19	35	0	45
4	0	0	55	162	0	11999	146	1639	19732	0	26818	22	39	0	39
5	0	0	755	35130	34781	9377	1040	820	34546	0	9932	10	56	0	35
6	0	0	62	164	0	12016	134	1620	20057	0	27354	22	39	0	39
7	24	0	75	160	2	12234	157	1328	21187	0	30571	26	48	0	26
8	18	0	114	24268	24111	7776	259	817	14738	0	15275	14	40	0	46
9	0	0	57	160	2	11871	142	1637	19936	0	27068	22	39	0	39
10	0	0	75	171	0	13295	150	1443	22625	0	31555	26	46	0	27
11	0	0	70	176	1	12067	143	1629	19945	0	27013	22	40	0	39
12	2	0	59	150	0	10660	129	929	17462	0	23315	19	35	0	46
13	27	0	73	162	1	12149	156	1296	21234	0	30775	26	48	0	25
14	0	0	63	150	1	10739	129	928	17889	0	23918	20	35	0	45
15	0	0	676	35086	34722	9280	1080	801	35062	0	10174	10	56	0	34

```
□
```

Step 1: Multiple Listeners

```
$ prstat -lm -c -n 20 5 6
PID USERNAME  UPR  SWG  TRP  TFL  DFL  LCK  SLP  LAT  VCX  ICX  SCL  SIG  LWPID  PROCESS/LMPNAME
17282 root      26,63 42,18 0,054 0,000 0,000 21,16 6,593 3,373 32483 1579 164K 0 10 dafileserver/rx_Listener
17282 root      26,21 41,72 0,056 0,000 0,000 22,09 6,579 3,354 32164 1440 163K 0 11 dafileserver/rx_Listener
17282 root      26,15 41,59 0,057 0,000 0,000 22,32 6,533 3,349 32094 1518 162K 0 8 dafileserver/rx_Listener
17282 root      26,08 41,14 0,062 0,000 0,000 22,99 6,437 3,291 31587 1608 160K 0 5 dafileserver/rx_Listener
17282 root      25,76 41,08 0,051 0,000 0,000 23,27 6,473 3,372 31592 1484 159K 0 3 dafileserver/rx_Listener
17282 root      25,52 40,42 0,060 0,000 0,000 24,39 6,343 3,271 31015 1449 157K 0 9 dafileserver/rx_Listener
17282 root      25,49 40,15 0,065 0,000 0,000 24,59 6,423 3,281 31017 1575 156K 0 7 dafileserver/rx_Listener
17282 root      25,32 40,12 0,061 0,000 0,000 24,73 6,462 3,310 31400 1582 155K 0 6 dafileserver/rx_Listener
17282 root      24,76 39,36 0,056 0,000 0,000 26,18 6,376 3,262 30601 1582 150K 0 4 dafileserver/rx_Listener
17282 root      24,77 39,14 0,075 0,000 0,000 26,54 6,262 3,217 30071 1699 150K 0 12 dafileserver/rx_Listener
17282 root      0,626 1,148 0,001 0,000 0,000 97,91 0,025 0,094 739 48 3395 0 108 dafileserver
17282 root      0,813 1,134 0,003 0,000 0,000 97,94 0,021 0,088 689 59 3356 0 101 dafileserver
17282 root      0,808 1,129 0,003 0,000 0,000 97,96 0,017 0,088 703 47 3449 0 113 dafileserver
17282 root      0,805 1,116 0,002 0,000 0,000 97,98 0,011 0,086 679 53 3313 0 36 dafileserver
17282 root      0,797 1,123 0,004 0,000 0,000 97,97 0,018 0,091 737 61 3296 0 95 dafileserver
17282 root      0,803 1,105 0,001 0,000 0,000 98,00 0,012 0,083 671 51 3288 0 79 dafileserver
17282 root      0,797 1,106 0,001 0,000 0,000 97,99 0,015 0,092 728 53 3428 0 85 dafileserver
17282 root      0,796 1,103 0,002 0,000 0,000 97,99 0,014 0,092 736 58 3265 0 125 dafileserver
17282 root      0,802 1,095 0,003 0,000 0,000 98,00 0,013 0,085 716 62 3337 0 116 dafileserver
17282 root      0,794 1,098 0,003 0,000 0,000 98,00 0,010 0,097 699 63 3279 0 48 dafileserver
Total: 239 processes, 1270 lups, load averages: 7.88, 6.42, 5.06
```

□

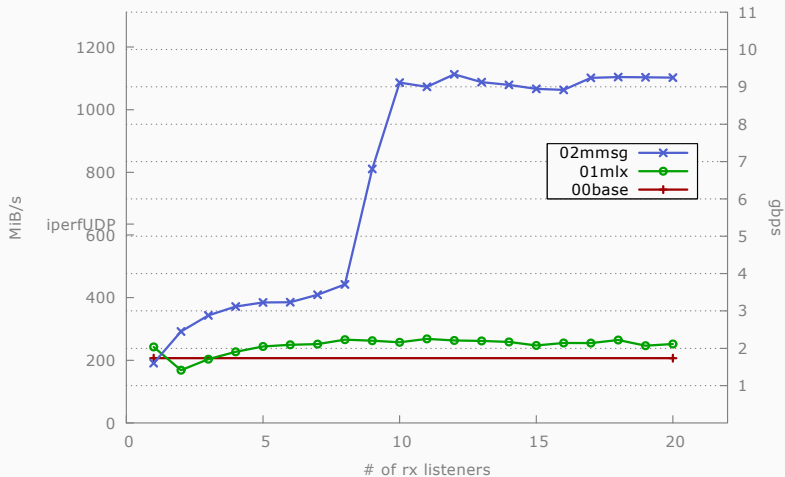
Syscall Overhead

- Each packet received is 1 syscall, plus locking
 - in `rx_Listener`
- Each packet sent is 1 syscall, plus locking
 - sometimes in `rx_Listener`
- We must send packets separately, but:

Step 2: recvmmsg/sendmmsg

- `recvmmsg()/sendmmsg()` (note the extra `m`)
 - Solaris 11.4+, RHEL 7+
- Receive same-call packets in bulk, `qsort()`
- Also benefits platforms without `*mmsg`

Step 2: recvmmsg/sendmmsg



Step 2: recvmmsg/sendmmsg

```
$ mpstat 5 6
```

CPU	minf	mjf	xcal	intr	ithr	csw	icsw	migr	smtx	srw	syscl	usr	sys	st	idl
0	0	0	856	1809	102	8324	1801	2324	111387	0	38403	17	79	0	4
1	0	0	813	1468	0	8142	1758	2188	106130	0	38177	17	79	0	4
2	0	0	702	1571	0	8507	1843	2339	110093	0	39689	18	78	0	4
3	0	0	710	1530	72	8055	1739	2169	105634	0	37317	17	79	0	4
4	0	0	668	1544	0	8402	1819	2352	111156	0	38241	17	79	0	4
5	0	0	5785	43032	42269	6271	1794	1192	77709	0	20890	10	88	0	2
6	0	0	754	1566	0	8563	1853	2387	111051	0	40139	18	78	0	4
7	0	0	698	1179	2	7467	1338	2359	103343	0	37033	17	79	0	3
8	0	0	685	1506	0	8304	1806	2229	109504	0	38898	18	78	0	4
9	0	0	667	1465	2	8180	1743	2282	107824	0	37073	17	79	0	4
10	0	0	711	1580	0	8442	1871	2263	109184	0	39748	18	78	0	4
11	0	0	676	1466	0	8072	1730	2272	105882	0	37513	17	79	0	4
12	0	0	710	1538	0	8378	1818	2224	108856	0	39679	18	78	0	4
13	0	0	717	1101	0	7314	1274	2247	104400	0	36131	17	80	0	3
14	0	0	700	1582	1	8546	1877	2244	110930	0	39673	18	78	0	4
15	0	0	5604	46215	45559	5655	1650	1047	75345	0	19157	10	89	0	2

```
□
```

Step 2: recvmmsg/sendmmsg

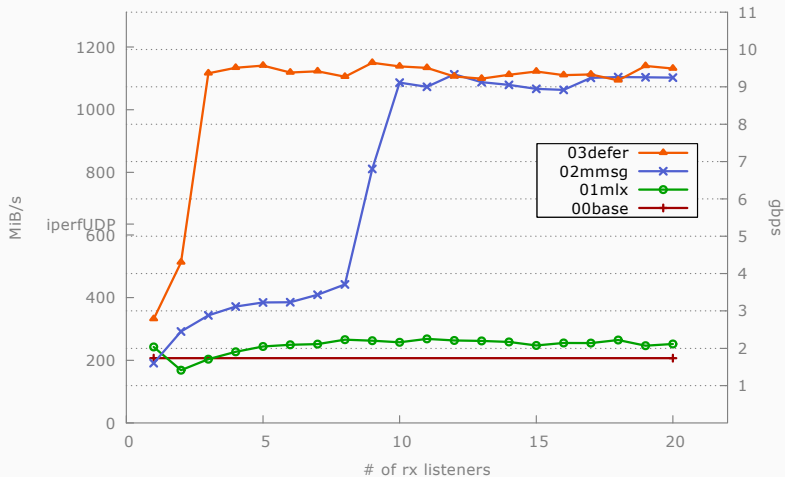
```
$ prstat -lm -c -n 20 5 6
PID USERNAME  U%   S%   TRP   TFL   DFL   LCK   SLP   LAT   VCX   ICX   SCL   SIG  LWPID  PROCESS/LMPNAME
27401 root      16,07 59,17 0,144 0,000 0,000 11,31 1,354 11,95 15171 9229 117K  0    12  dafileserver/rx_Listener
27401 root      16,16 58,78 0,150 0,000 0,000 11,52 1,403 11,98 15411 9112 117K  0    4  dafileserver/rx_Listener
27401 root      15,92 58,53 0,142 0,000 0,000 11,48 1,351 12,58 15242 9467 117K  0   10  dafileserver/rx_Listener
27401 root      16,00 58,21 0,152 0,000 0,000 11,45 1,352 12,84 15489 9448 116K  0    5  dafileserver/rx_Listener
27401 root      15,98 58,15 0,148 0,000 0,000 11,31 1,348 13,06 15294 9385 117K  0    6  dafileserver/rx_Listener
27401 root      15,95 57,81 0,149 0,000 0,000 11,58 1,394 13,12 15431 9432 117K  0    8  dafileserver/rx_Listener
27401 root      15,81 57,78 0,154 0,000 0,000 11,44 1,371 13,44 15252 9657 116K  0    3  dafileserver/rx_Listener
27401 root      15,79 57,39 0,160 0,000 0,000 11,95 1,344 13,37 15308 9578 114K  0   11  dafileserver/rx_Listener
27401 root      15,85 57,22 0,154 0,000 0,000 11,93 1,431 13,42 15492 9653 116K  0    9  dafileserver/rx_Listener
27401 root      15,69 56,01 0,168 0,000 0,000 11,71 1,356 15,07 15311 10189 112K  0    7  dafileserver/rx_Listener
27401 root      1,301 3,833 0,011 0,000 0,000 93,80 0,075 0,963 1713 487 19646 0   32  dafileserver
27401 root      1,300 3,803 0,009 0,000 0,000 93,80 0,079 1,012 1763 459 19629 0   52  dafileserver
27401 root      1,297 3,806 0,010 0,000 0,000 93,72 0,076 1,089 1667 487 19091 0   42  dafileserver
27401 root      1,296 3,785 0,008 0,000 0,000 93,88 0,068 0,959 1676 416 19276 0   22  dafileserver
27401 root      1,223 3,772 0,008 0,000 0,000 93,97 0,076 0,954 1596 471 18403 0  106  dafileserver
27401 root      1,275 3,710 0,008 0,000 0,000 93,99 0,077 0,943 1721 430 19077 0   77  dafileserver
27401 root      1,251 3,679 0,010 0,000 0,000 93,79 0,070 1,201 1614 586 18488 0   95  dafileserver
27401 root      1,262 3,643 0,011 0,000 0,000 93,90 0,084 1,096 1736 546 18774 0   74  dafileserver
27401 root      1,227 3,660 0,008 0,000 0,000 93,96 0,076 1,065 1624 468 18276 0   93  dafileserver
27401 root      1,213 3,667 0,009 0,000 0,000 94,01 0,085 1,019 1619 452 18249 0   30  dafileserver
Total: 237 processes, 1268 lups, load averages: 13.58, 11.17, 9.25
█
```


- Normally: buffer, then `sendmsg()`
- If the tx window is full:
 - Wait?
 - Overfill tx window
 - The listener calls `sendmsg()`
- Why?
 - Reduces context switching for LWP
 - Allows RPC threads to move on

Step 3: rxi_Start Defer

- Skip calling `rxi_Start()` in the listener
- Flag call instead
- Wakeup `rx_Write`, which calls `rxi_Start()`
- Only when `rx_Write` is waiting for the tx window
- Alternate approach: process packets in `rx_Write`

Step 3: rxi_Start Defer



Step 3: rxi_Start Defer

```
$ mpstat 5 6
CPU minf mjf xcal intr ithr csw icsw migr smtx srw syscl usr sys st idl
0 0 0 2377 3878 102 14607 4227 3138 113594 0 26160 15 85 0 0
1 0 0 2267 3767 0 15363 4463 3245 122662 0 26876 15 84 0 0
2 0 0 2203 3499 0 14534 4154 3129 115423 0 26003 15 84 0 1
3 0 0 2367 3801 49 15302 4473 3192 123713 0 27241 16 84 0 1
4 0 0 2197 3551 0 14624 4181 3188 117099 0 25632 15 85 0 0
5 0 0 10512 50634 48967 9121 2724 1770 82633 0 13869 9 91 0 0
6 0 0 2295 3636 0 14905 4288 3185 116756 0 26585 15 84 0 0
7 0 0 2578 3551 2 15369 4156 3570 114341 0 27981 17 83 0 0
8 0 0 2265 3474 0 14418 4118 3085 114486 0 25436 15 85 0 0
9 0 0 2229 3713 3 15282 4426 3241 121258 0 27257 16 83 0 1
10 0 0 2285 3582 0 14696 4253 3149 113719 0 25899 15 84 0 1
11 0 0 2255 3655 0 15072 4318 3259 124190 0 26748 16 84 0 1
12 0 0 2260 3614 0 14822 4245 3157 116169 0 26271 15 84 0 0
13 7 0 2517 3489 0 15132 4105 3581 115979 0 27502 17 83 0 1
14 0 0 2235 3639 1 14820 4300 3160 117537 0 26079 15 84 0 1
15 0 0 10472 51229 49631 8721 2635 1727 80970 0 13163 8 91 0 0
```


Step 3: rxi_Start Defer

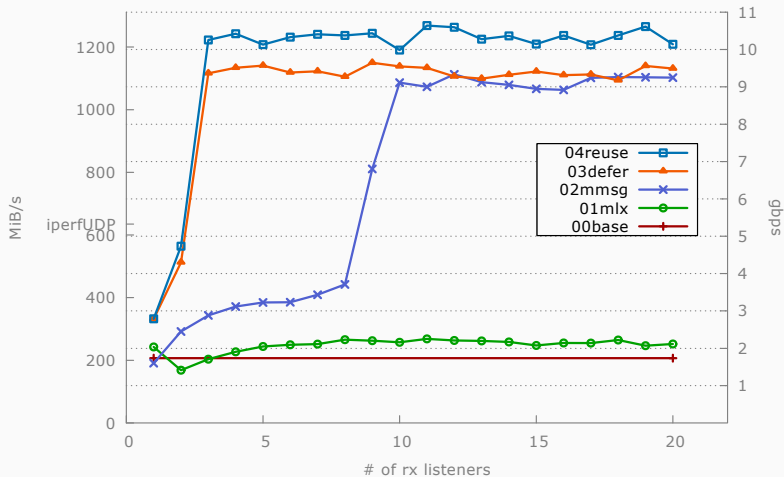
```
$ prstat -lm -c -n 20 5 6
PID USERNAME  U%   S%   TRP  TFL  DFL  LCK  SLP  LAT  VCX  ICX  SCL  SIG  LWPID  PROCESS/LMPNAME
7638 root      12.76 13.40 0.275 0.000 0.000 22.65 6.163 44.75 33046 13944 82886 0 6  dafileservr/rx_Listener
7638 root      12.68 13.35 0.273 0.000 0.000 23.23 6.186 44.28 33689 13772 83816 0 5  dafileservr/rx_Listener
7638 root      11.69 12.74 0.286 0.000 0.000 21.58 5.932 47.77 31642 14378 77395 0 12 dafileservr/rx_Listener
7638 root      11.86 12.50 0.282 0.000 0.000 21.34 5.870 48.14 31666 14380 78085 0 4  dafileservr/rx_Listener
7638 root      11.55 12.29 0.281 0.000 0.000 21.69 5.988 48.20 31792 13973 76394 0 9  dafileservr/rx_Listener
7638 root      11.24 12.24 0.280 0.000 0.000 22.39 5.782 48.07 31672 14374 74723 0 3  dafileservr/rx_Listener
7638 root      10.96 11.73 0.296 0.000 0.000 22.20 5.772 49.04 31030 14941 73463 0 7  dafileservr/rx_Listener
7638 root      10.81 11.80 0.291 0.000 0.000 20.73 5.738 50.63 30650 14525 72304 0 10 dafileservr/rx_Listener
7638 root      10.90 11.68 0.280 0.000 0.000 22.06 5.861 49.22 30653 14287 72958 0 11 dafileservr/rx_Listener
7638 root      10.49 11.27 0.299 0.000 0.000 21.17 5.550 51.22 30276 15082 69825 0 8  dafileservr/rx_Listener
7638 root      1.402 3.364 0.034 0.000 0.000 77.49 0.212 11.49 5535 1755 13404 0 60 dafileservr
7638 root      1.427 9.306 0.033 0.000 0.000 78.32 0.199 10.71 5597 1751 13664 0 67 dafileservr
7638 root      1.417 9.085 0.031 0.000 0.000 79.33 0.203 9.933 5721 1680 13892 0 30 dafileservr
7638 root      1.375 9.053 0.034 0.000 0.000 77.19 0.210 12.14 5442 1920 13202 0 29 dafileservr
7638 root      1.378 9.048 0.034 0.000 0.000 75.62 0.200 13.72 5545 2125 13539 0 80 dafileservr
7638 root      1.373 9.033 0.034 0.000 0.000 78.19 0.201 11.17 5531 1939 13328 0 105 dafileservr
7638 root      1.404 8.978 0.031 0.000 0.000 80.23 0.198 9.160 5631 1591 13742 0 70 dafileservr
7638 root      1.357 8.994 0.036 0.000 0.000 77.54 0.216 11.86 5373 2055 13041 0 48 dafileservr
7638 root      1.374 8.926 0.039 0.000 0.000 75.37 0.206 14.09 5428 2256 13235 0 126 dafileservr
7638 root      1.355 8.923 0.031 0.000 0.000 78.31 0.209 11.17 5407 1826 13192 0 87 dafileservr
Total: 236 processes, 1267 lups, load averages: 25.51, 22.51, 18.74
```

□

recvmsg() parallelization

- Remember: `recvmsg()` itself internally serialized
 - *per socket*
- `SO_REUSEPORT` allows for sockets on same addr
 - Solaris 11+, RHEL6.5+
- Packets assigned to sockets based on configurable hash
 - Default: IP and port for source and destination

Step 4: SO_REUSEPORT



Step 4: SO_REUSEPORT

```
$ mpstat 5 6
CPU minf mjf xcal intr ithr csw icsw migr smtx srw syscl usr sys st idl
0 0 0 1225 3656 103 11575 3639 2338 111583 0 20132 16 83 0 0
1 0 0 5762 26852 24673 8814 2700 1778 83271 0 13600 12 88 0 0
2 0 0 1083 3467 1 11951 3736 2353 114350 0 20898 17 83 0 0
3 0 0 1063 3419 76 11445 3579 2321 108043 0 19541 16 84 0 0
4 0 0 1052 3404 1 11808 3698 2336 112030 0 20455 17 83 0 0
5 0 0 11727 46396 44545 7796 2480 1665 78798 0 11820 10 89 0 0
6 0 0 1086 3416 1 11886 3720 2365 109512 0 20547 17 83 0 0
7 0 0 6059 26717 24804 8507 2438 1844 80091 0 13130 11 89 0 0
8 0 0 1058 3366 1 11716 3628 2332 111140 0 20569 17 83 0 0
9 0 0 1131 3165 3 11296 3371 2463 104376 0 19317 16 84 0 0
10 0 0 1067 3479 1 11992 3741 2340 112746 0 20649 17 83 0 0
11 0 0 1080 3334 1 11493 3550 2337 109345 0 19698 16 84 0 0
12 0 0 1022 3382 1 11769 3661 2332 109851 0 20470 17 83 0 0
13 0 0 1094 3025 1 11390 3272 2529 105696 0 20369 17 83 0 0
14 0 0 1055 3320 2 11490 3570 2332 110170 0 19883 16 84 0 0
15 0 0 10838 46063 44177 7925 2374 1833 79821 0 12135 11 89 0 0
```

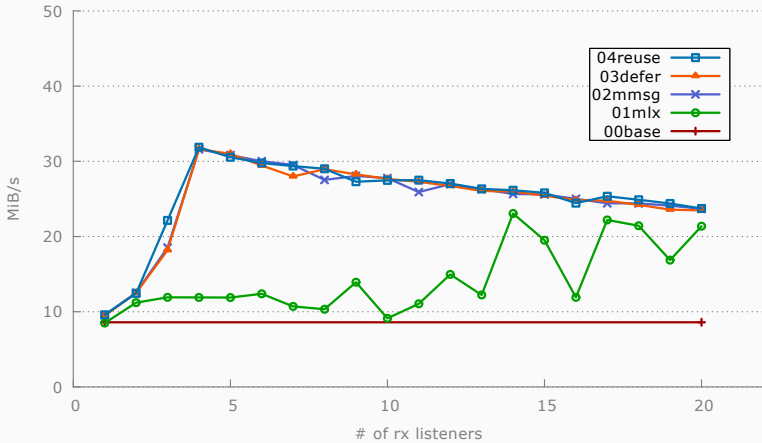
Step 4: SO_REUSEPORT

```
$ prstat -lm -c -n 20 5 6
  PID USERNAME  UPR  SWG  TRP  TFL  DFL  LCK  SLP  LAT  VCX  ICX  SCL  SIG  LWPID  PROCESS/LMPNAME
17763 root      14,79 8,753 0,349 0,000 0,000 11,37 5,750 58,99 14998 12478 58916 0 9  dafileservr/rx_Listener
17763 root      14,51 8,770 0,349 0,000 0,000 11,44 2,250 62,68 13625 12436 54307 0 3  dafileservr/rx_Listener
17763 root      14,70 8,462 0,334 0,000 0,000 12,23 2,579 61,70 13157 11893 53258 0 4  dafileservr/rx_Listener
17763 root      14,68 8,471 0,339 0,000 0,000 11,46 2,758 62,30 13293 12235 52402 0 10 dafileservr/rx_Listener
17763 root      14,65 8,490 0,345 0,000 0,000 11,78 2,966 61,77 13588 12380 54119 0 11 dafileservr/rx_Listener
17763 root      14,47 8,522 0,343 0,000 0,000 11,44 6,931 58,29 15585 12653 59340 0 5  dafileservr/rx_Listener
17763 root      14,20 8,273 0,331 0,000 0,000 12,02 5,283 59,90 14435 12095 55992 0 8  dafileservr/rx_Listener
17763 root      13,62 8,153 0,337 0,000 0,000 11,49 10,02 56,38 16946 12690 60754 0 6  dafileservr/rx_Listener
17763 root      13,65 8,011 0,348 0,000 0,000 11,51 5,924 60,56 14860 12915 55245 0 7  dafileservr/rx_Listener
17763 root      12,88 7,578 0,340 0,000 0,000 11,38 5,835 61,99 14466 12398 52527 0 12 dafileservr/rx_Listener
17763 root      1,298 10,35 0,034 0,000 0,000 76,53 0,262 11,36 4217 1636 10181 0 90 dafileservr
17763 root      1,219 10,35 0,025 0,000 0,000 79,13 0,249 9,030 4203 1247 10019 0 63 dafileservr
17763 root      1,215 10,22 0,029 0,000 0,000 78,96 0,250 9,323 4095 1339 9841 0 133 dafileservr
17763 root      1,223 10,17 0,032 0,000 0,000 77,37 0,253 10,95 4114 1601 10048 0 122 dafileservr
17763 root      1,209 10,18 0,029 0,000 0,000 77,13 0,235 11,21 4088 1514 9909 0 88 dafileservr
17763 root      1,227 10,11 0,038 0,000 0,000 75,48 0,241 12,91 4135 1721 10021 0 130 dafileservr
17763 root      1,203 10,12 0,026 0,000 0,000 79,12 0,227 9,311 4008 1281 9863 0 59 dafileservr
17763 root      1,200 10,10 0,034 0,000 0,000 75,70 0,250 12,72 3989 1671 9693 0 138 dafileservr
17763 root      1,209 10,07 0,032 0,000 0,000 78,21 0,220 10,26 4004 1540 9940 0 57 dafileservr
17763 root      1,224 10,02 0,028 0,000 0,000 79,43 0,213 9,082 4122 1388 10224 0 86 dafileservr
Total: 237 processes, 1268 lups, load averages: 25,84, 23,27, 23,15
```

□

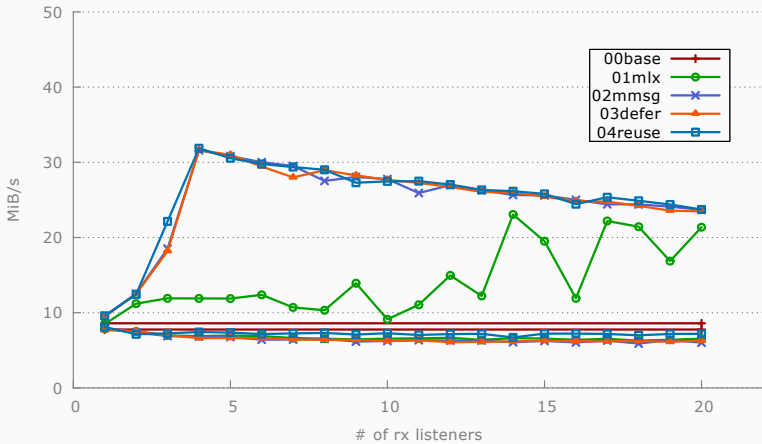
Small RPCs

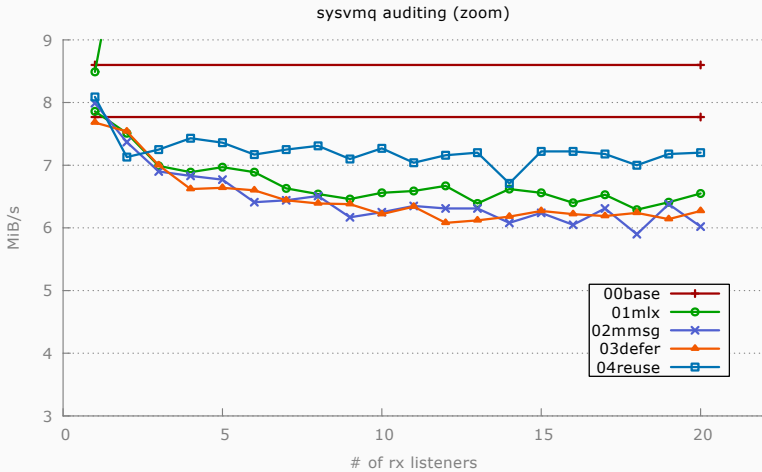
Step 4: SO_REUSEPORT (small)



- So far, default options besides `-p`
- What options matter?
- `-auditlog`

sysvmq auditing





- Audit subsystem uses one big global lock
- Addressed for new pipe audit interface
- See “OpenAFS Audit Interfaces enhancements” tomorrow

Lessons Learned

- Recording per-function runtimes is way too heavyweight
 - DTrace profile probes vs pid
- *Must* verify profiling performance impact
- Test, don't assume
 - VMs
 - localhost
 - auditlog

Future Possibilities

- More efficient ACK processing?
- Revisit jumbograms
- AF_RXRPC
- Kernel client improvements
- TCP (DPF)

Top commit

<https://gerrit.openafs.org/13621>

Public

<https://gerrit.openafs.org/#/q/topic:recvmmsg>

<https://gerrit.openafs.org/#/q/topic:sendmmsg>

Drafts

<https://gerrit.openafs.org/#/q/topic:multi-listener>

https://gerrit.openafs.org/#/q/topic:rxi_startdefer

<https://gerrit.openafs.org/#/q/topic:reuseport>

Slides

<http://dson.org/talks>

?