ceph

Ceph中国社区 | IT大咖说 知识共享平台

CEPHALOCON APAC 2018
THE FUTURE OF STORAGE
22-23 March 2018 | BEIJING

Reconsidering tracing in Ceph

Mohamad Gebai
Software Engineer at SUSE
@mogeb

Outline

- Vector vs list vs deque in bufferlist
- push_back()
- to_str()
- push_front()
- Tracing demo

Hypothesis

- Will changing the data structure behind bufferlist help performance?

- Will moving from a std::list to a std::vector make bufferlist more cache-friendly?

- rados bench isn’t targeted enough, improvement is too little to notice

- Solution: why not use tracing?

Code instrumentation

```
void buffer::list::push_front(ptr& bp) {
  if (bp.length() == 0)
    return;
  tracepoint(analyze, bufferlist_push_front, ENTER);
  _buffers.insert(_buffers.begin(), bp);
  tracepoint(analyze, bufferlist_push_front, EXIT);

  _len += bp.length();
}
```

Code instrumentatio

- We can add PMU counters in the context of each tracepoint

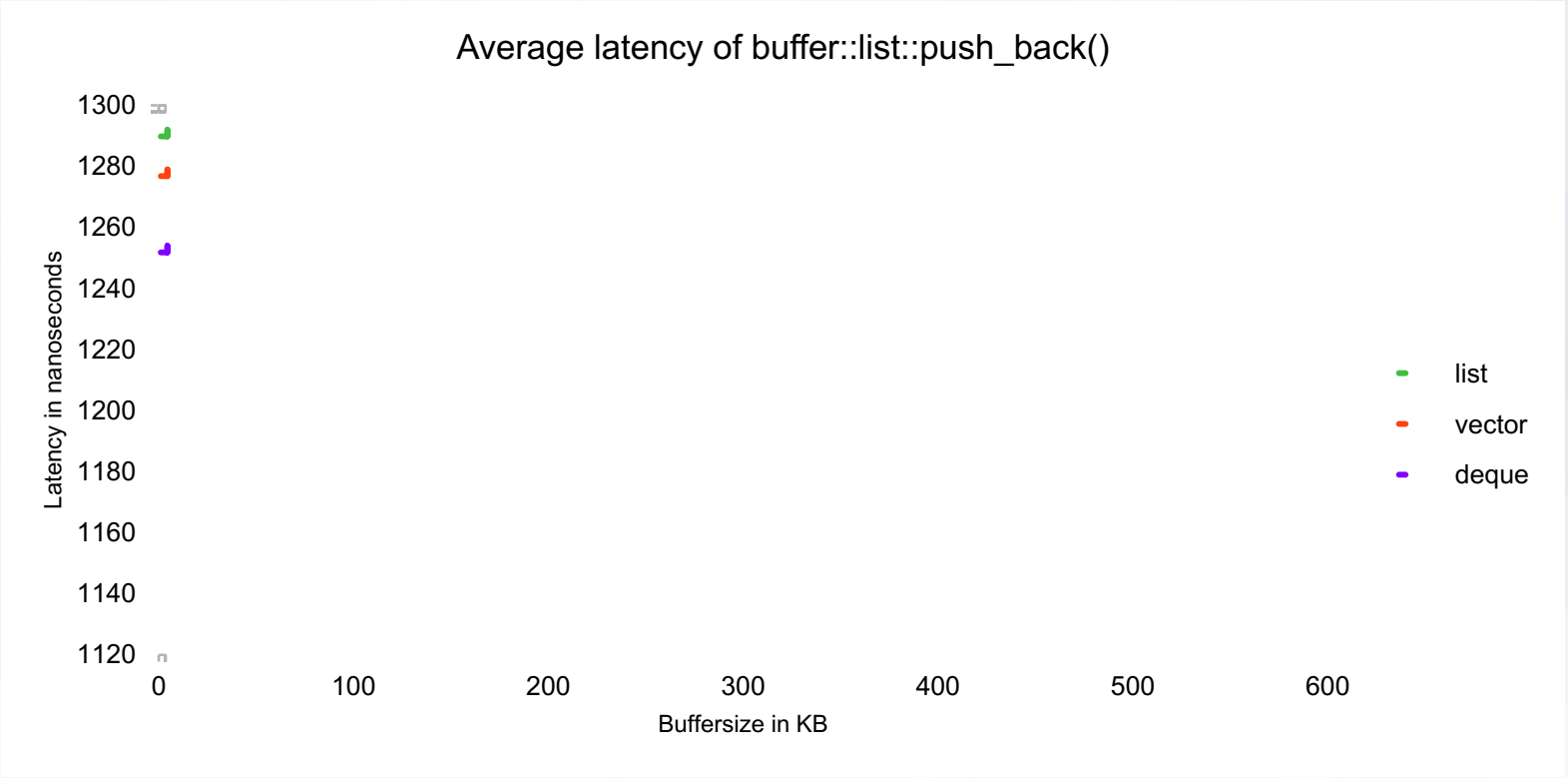
```
$> lttng create -o .
[...]
$> lttng add-context -u -t perf:thread:cache-misses -t perf:thread:dTLB-load-misses -t
perf:thread:L1-dcache-load-misses -t perf:thread:instructions
```

- Instructions can let us guess which code paths are taken (slow path vs fast path)

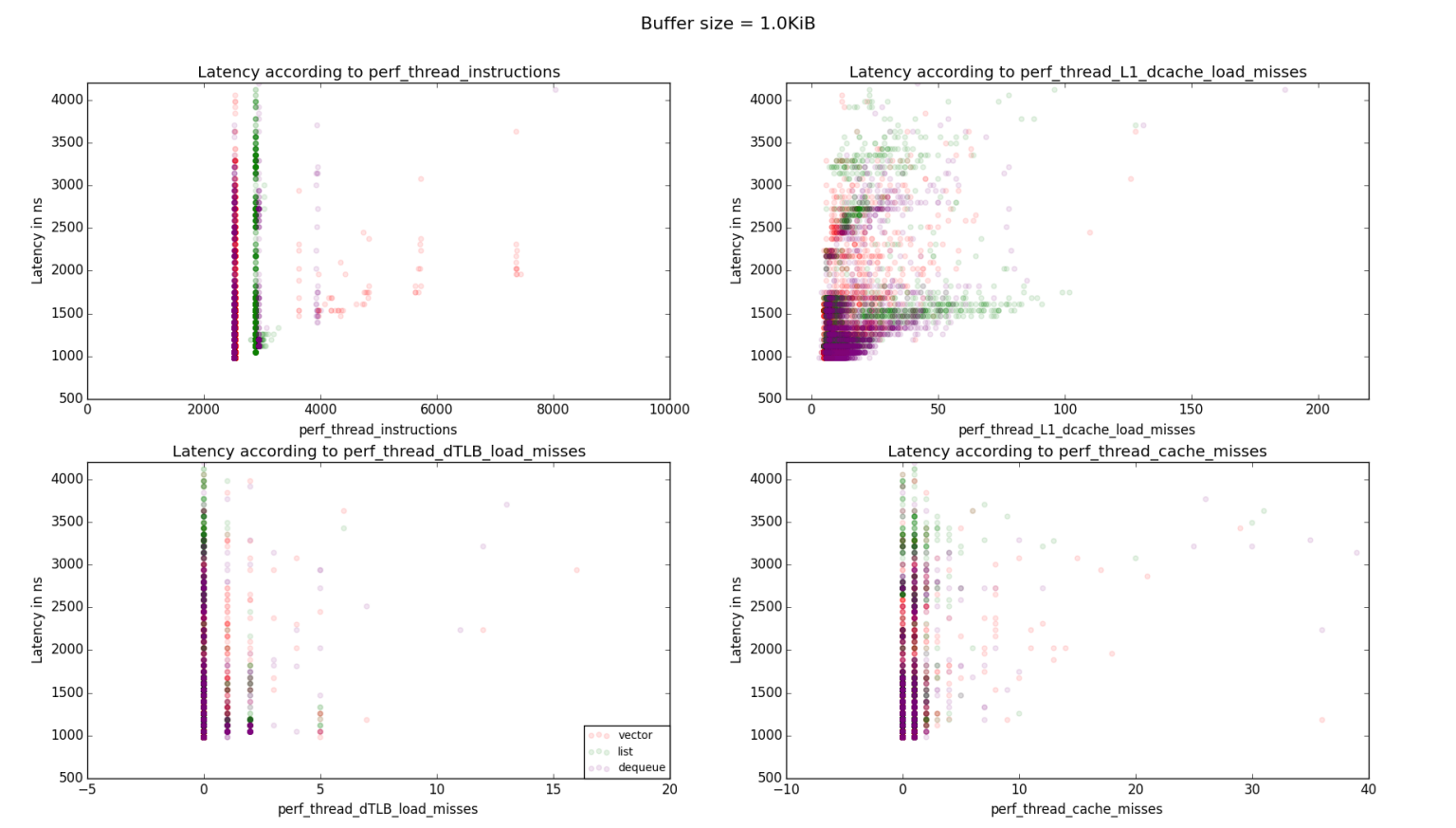
Access latency

- 1000 bufferlists
- 1000 buffer per bufferlist
- Vary buffer size
- Run push_back(), push_front() and to_str() workloads
- L1 cache size: 32 KB
- L2 cache size: 256 KB
- **Microbenchmarks in unit tests**

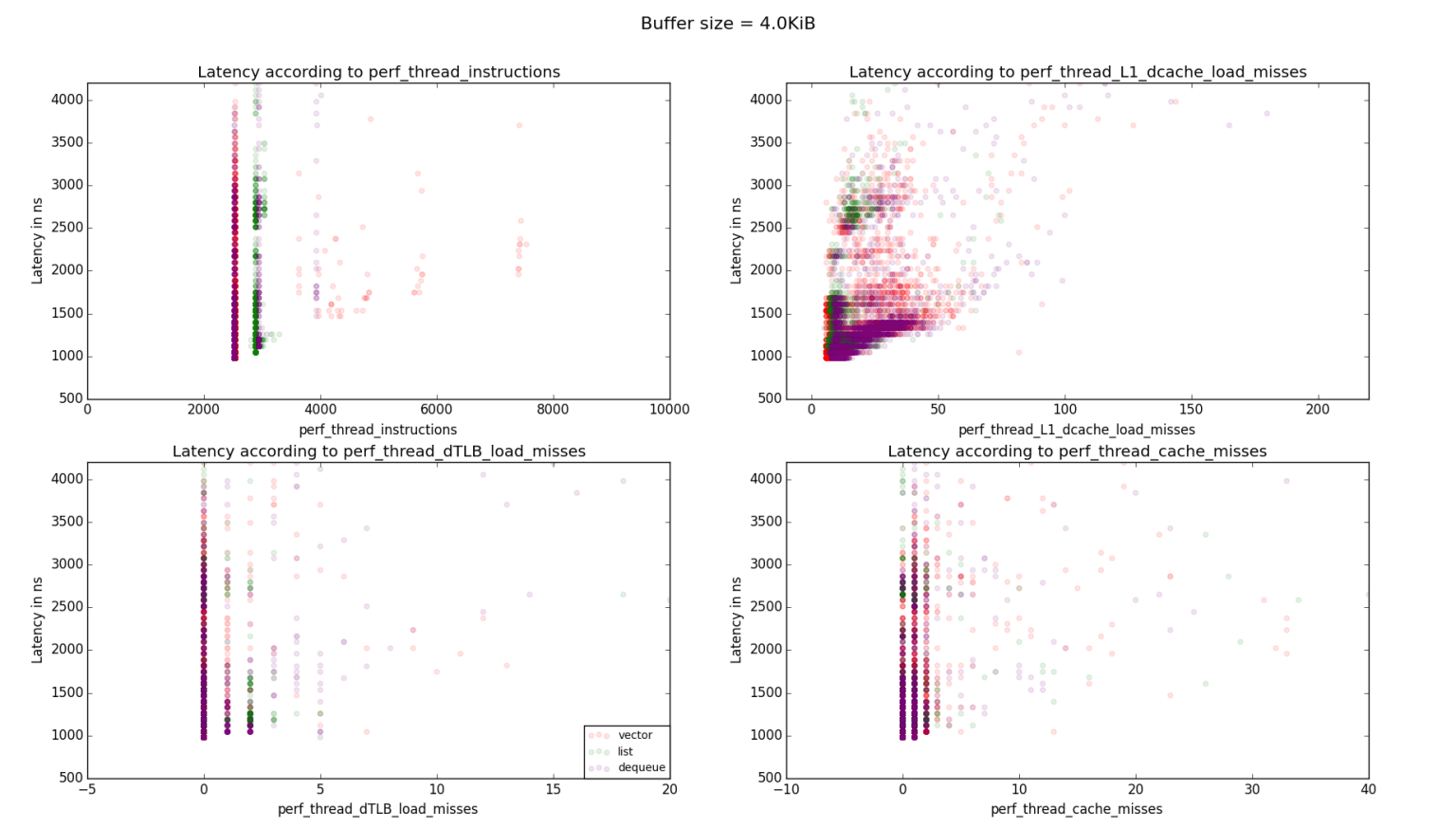
push_back() latency



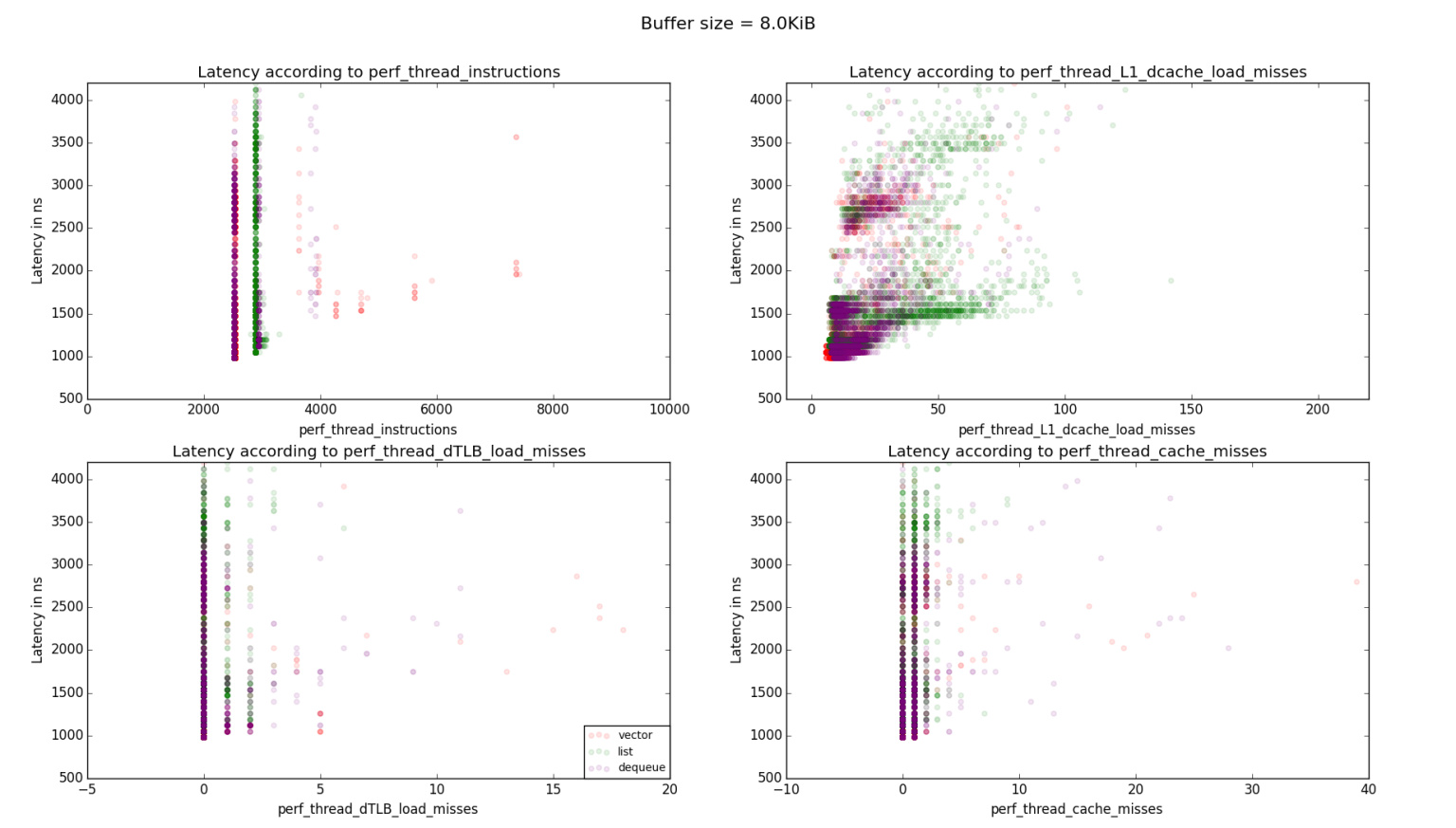
push_back() latency



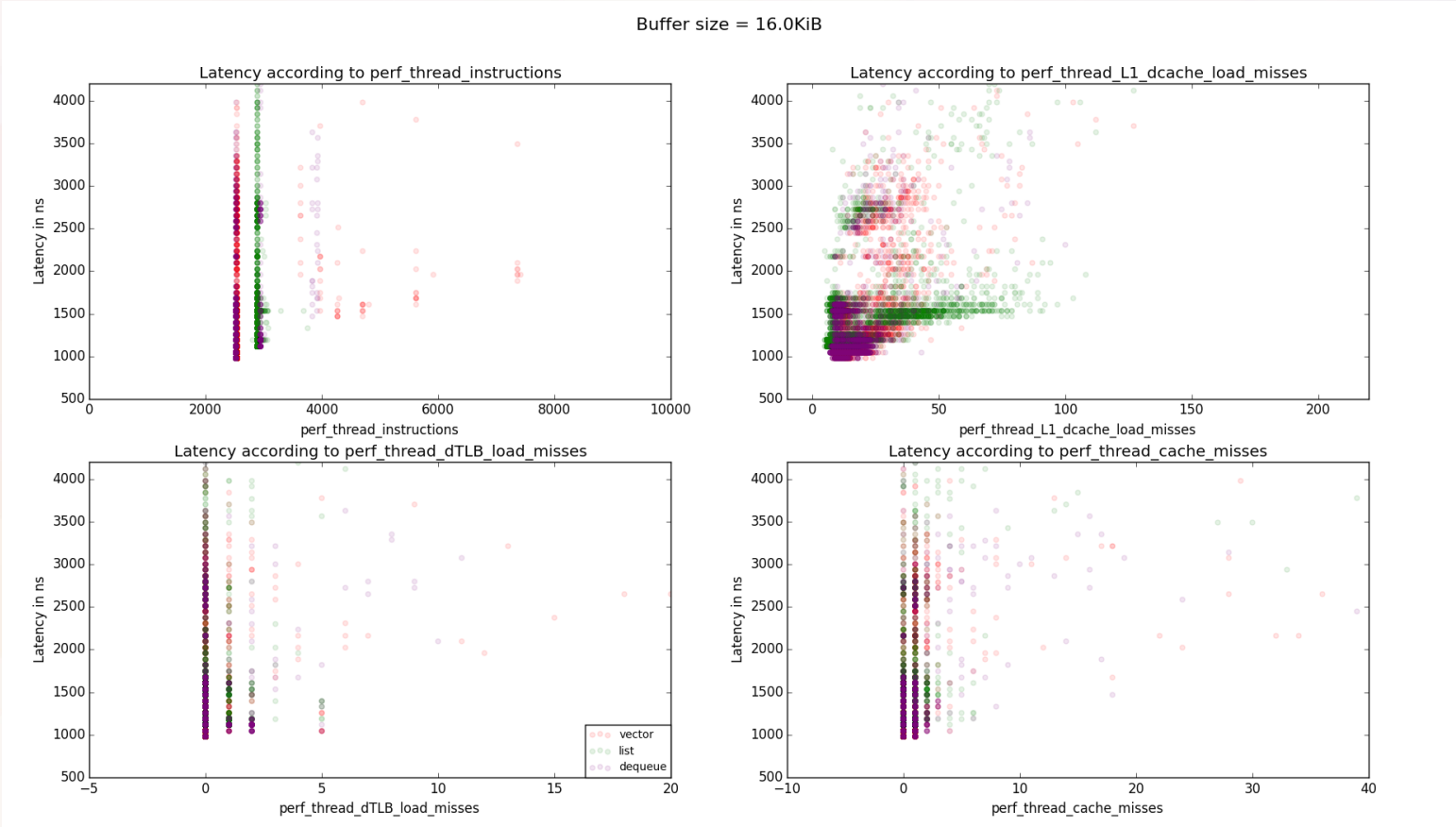
push_back() latency



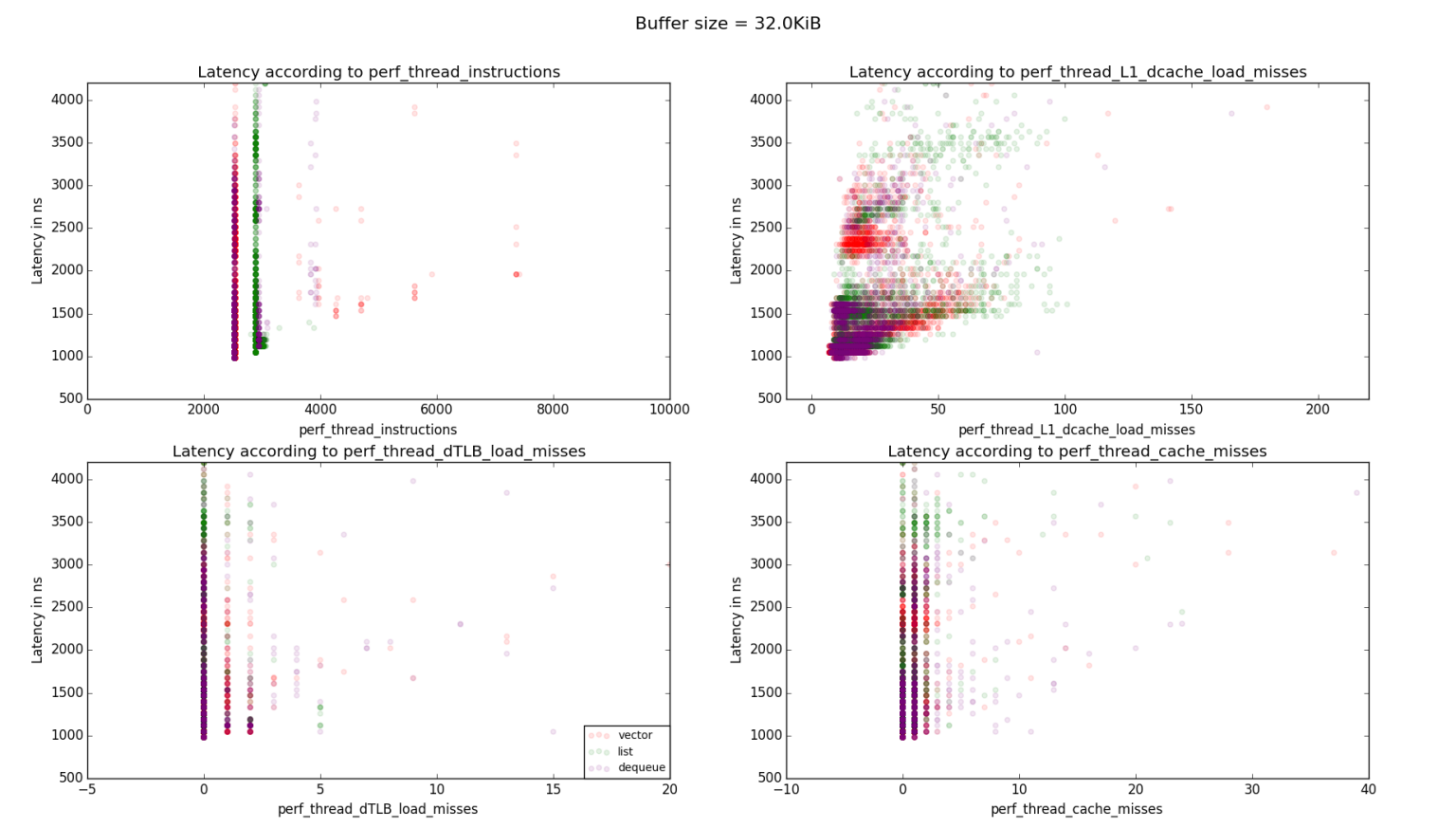
push_back() latency



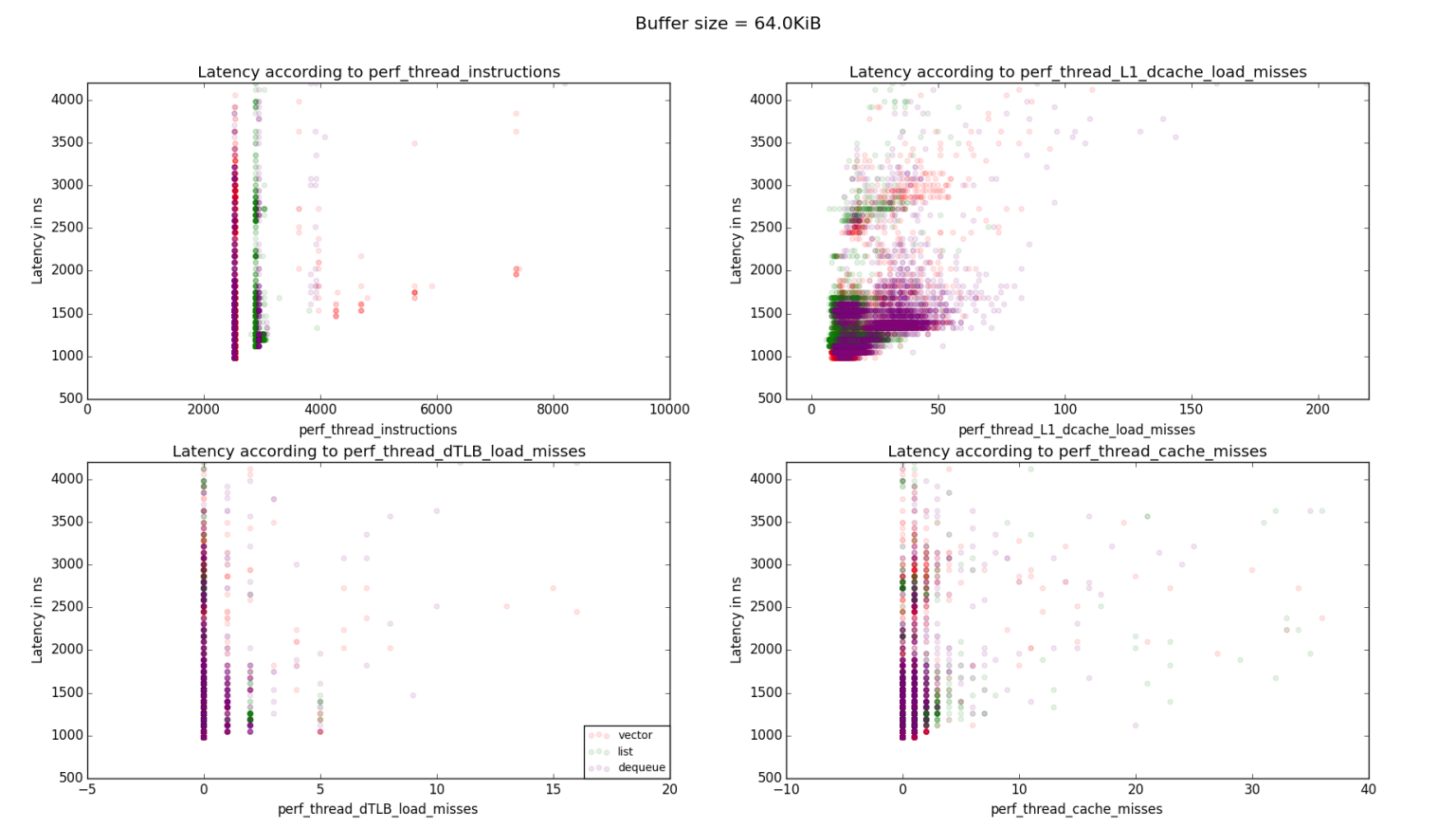
push_back() latency



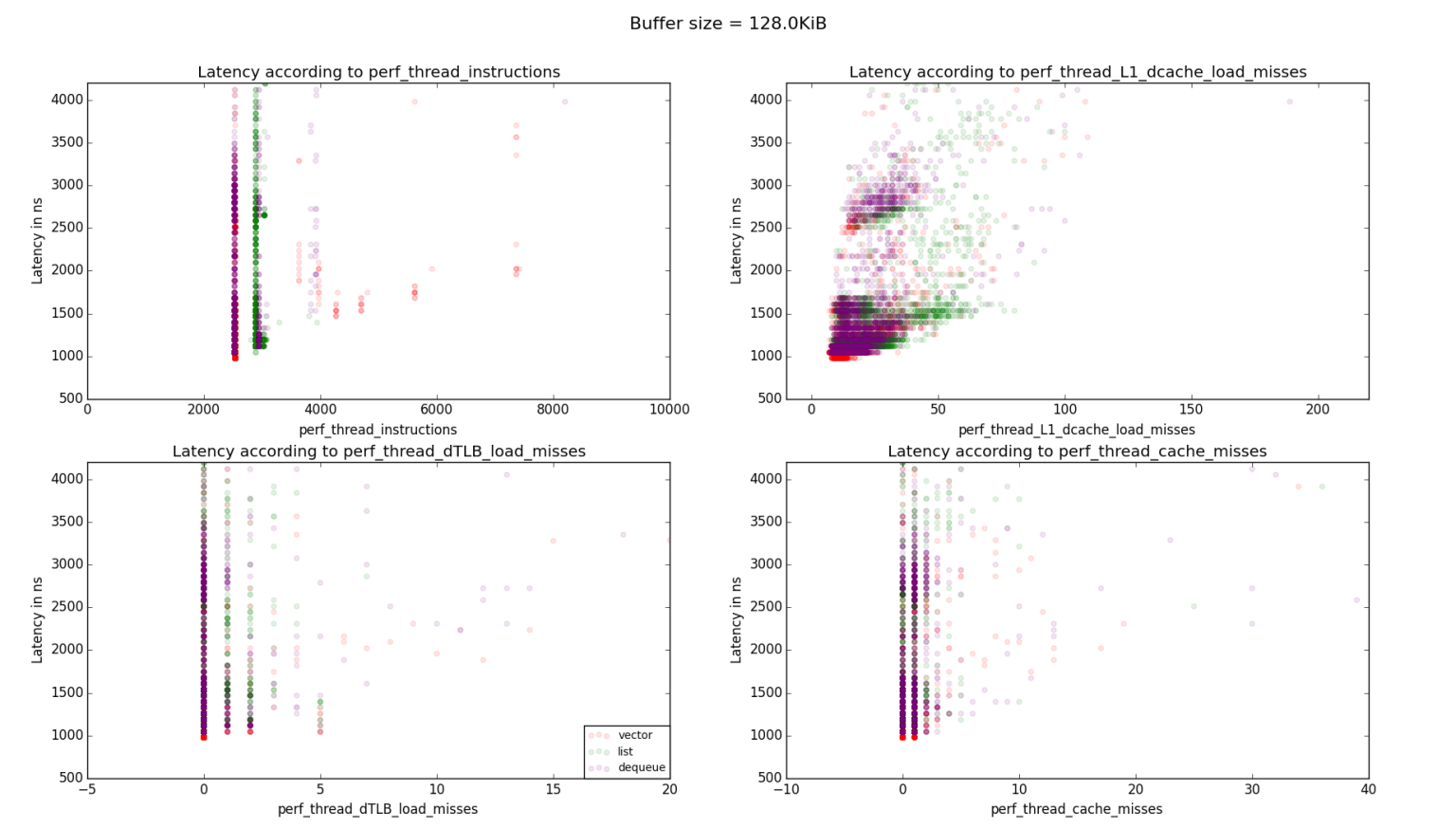
push_back() latency



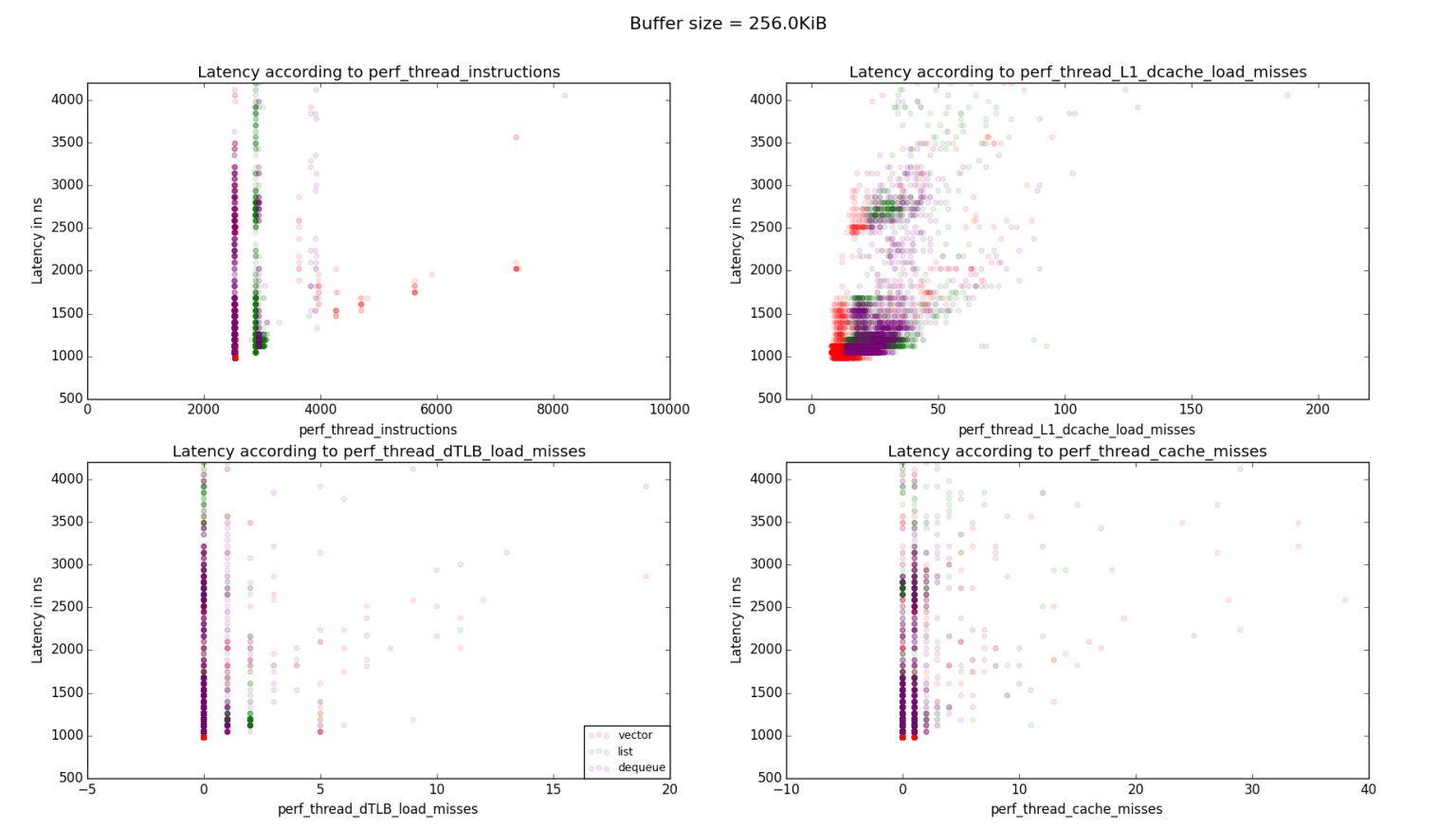
push_back() latency



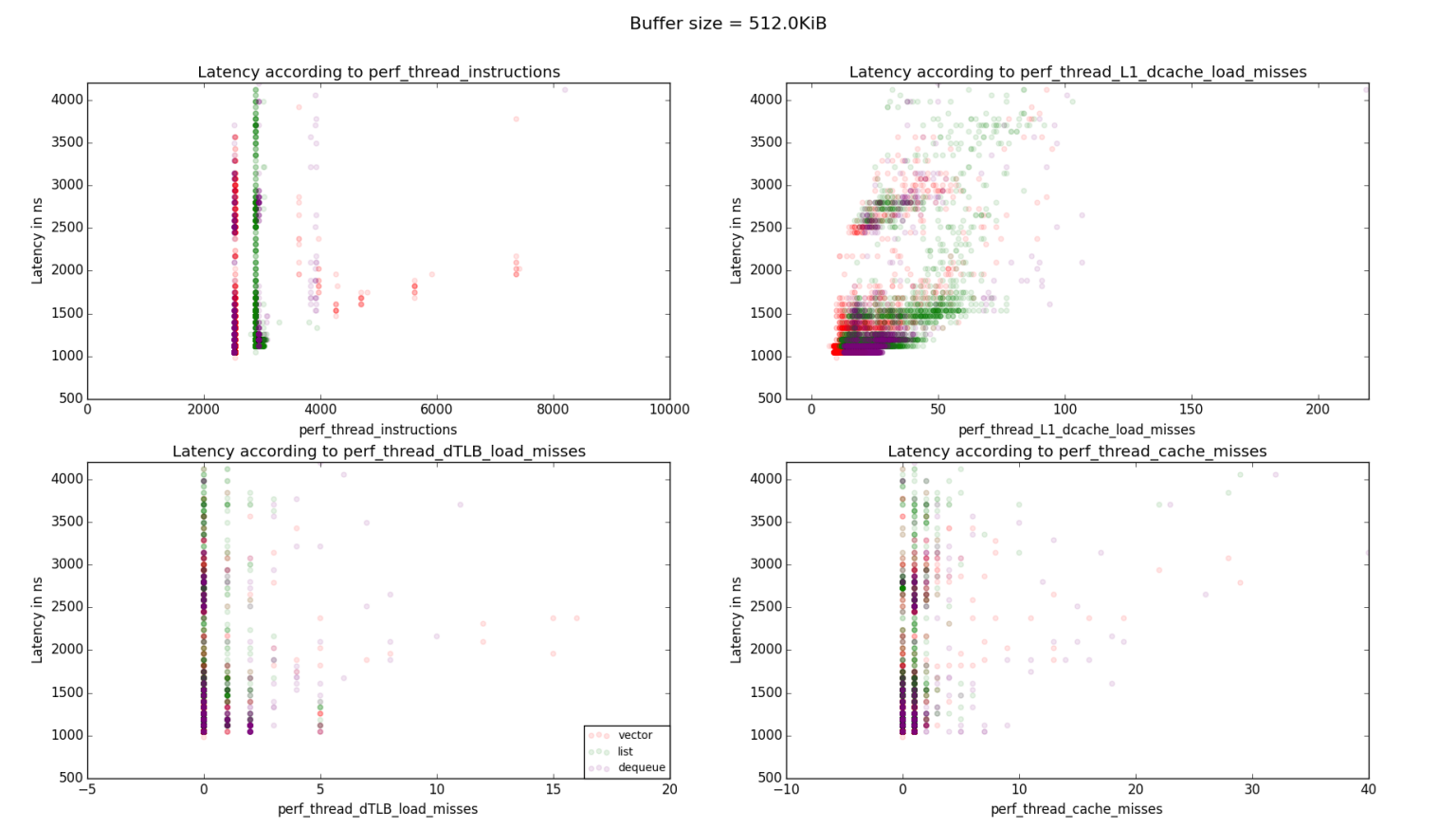
push_back() latency



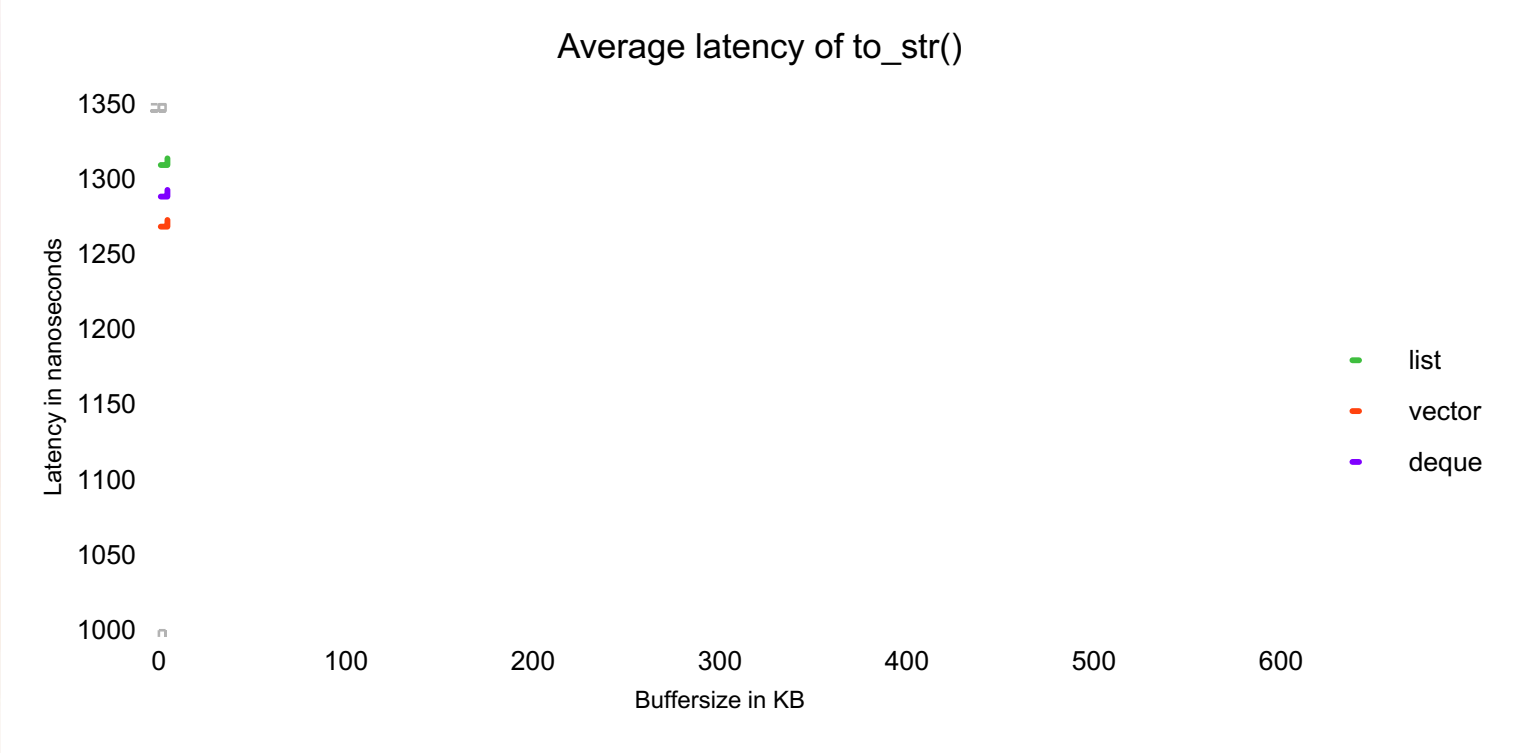
push_back() latency



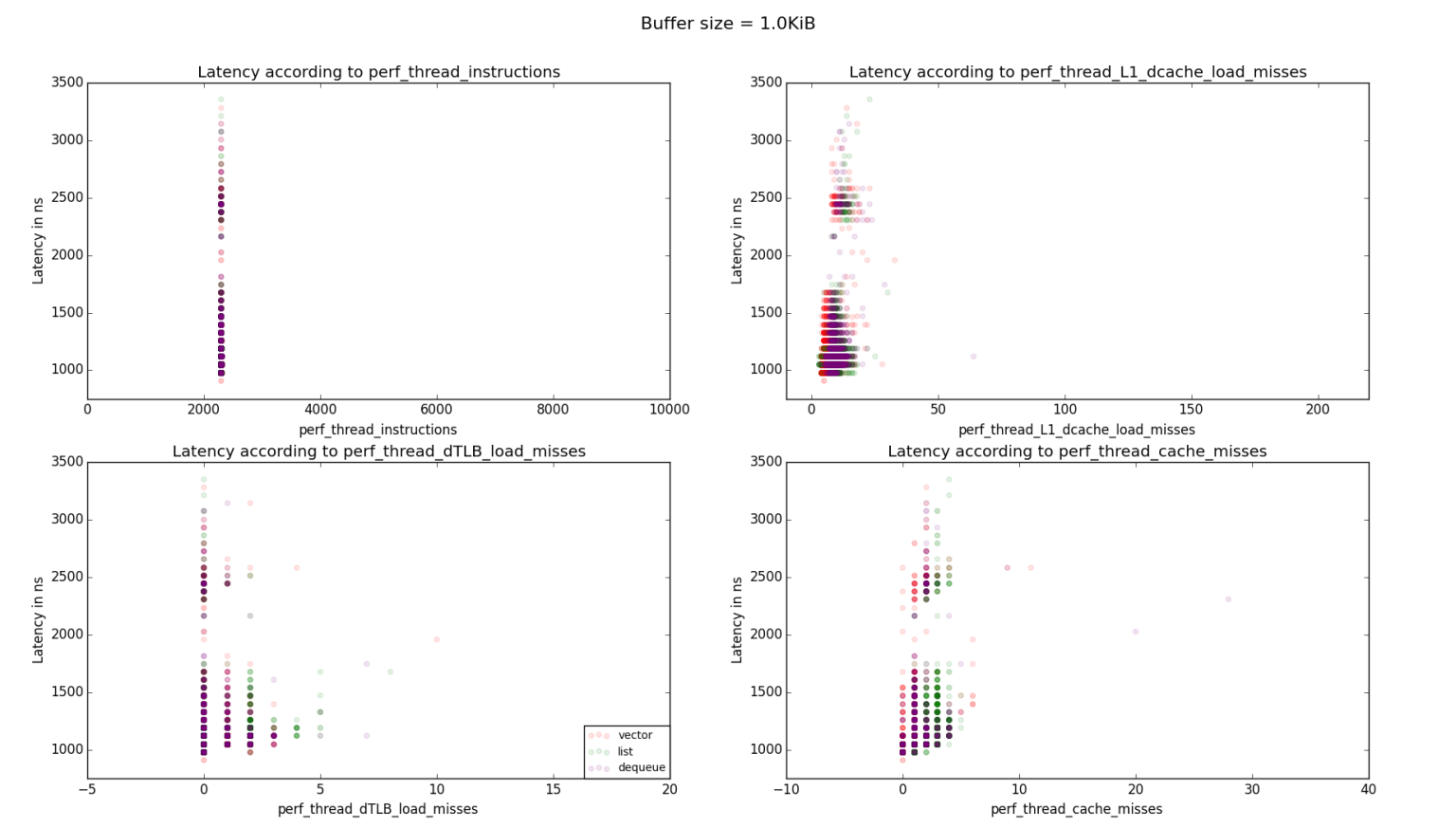
push_back() latency



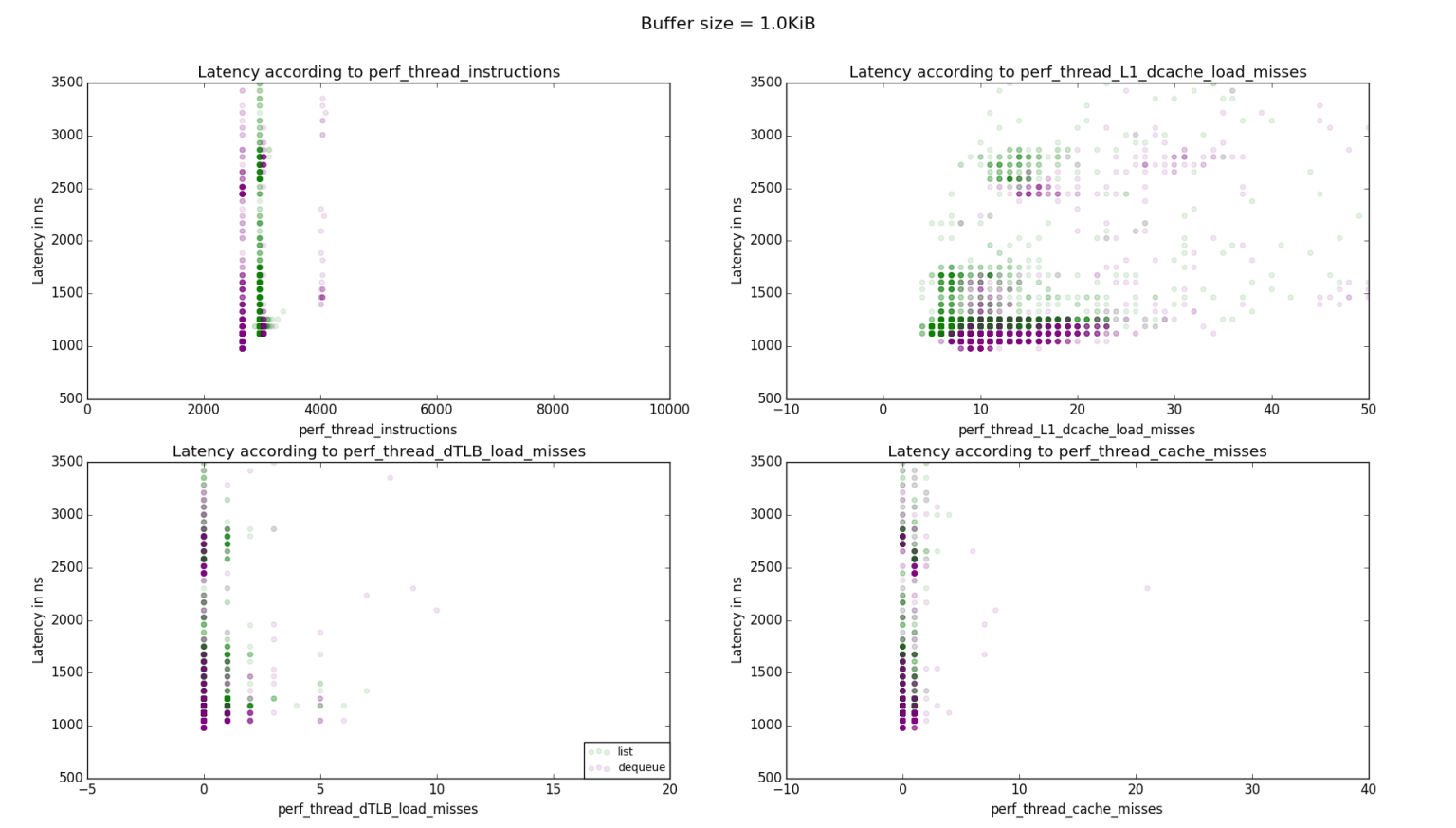
To str() latency



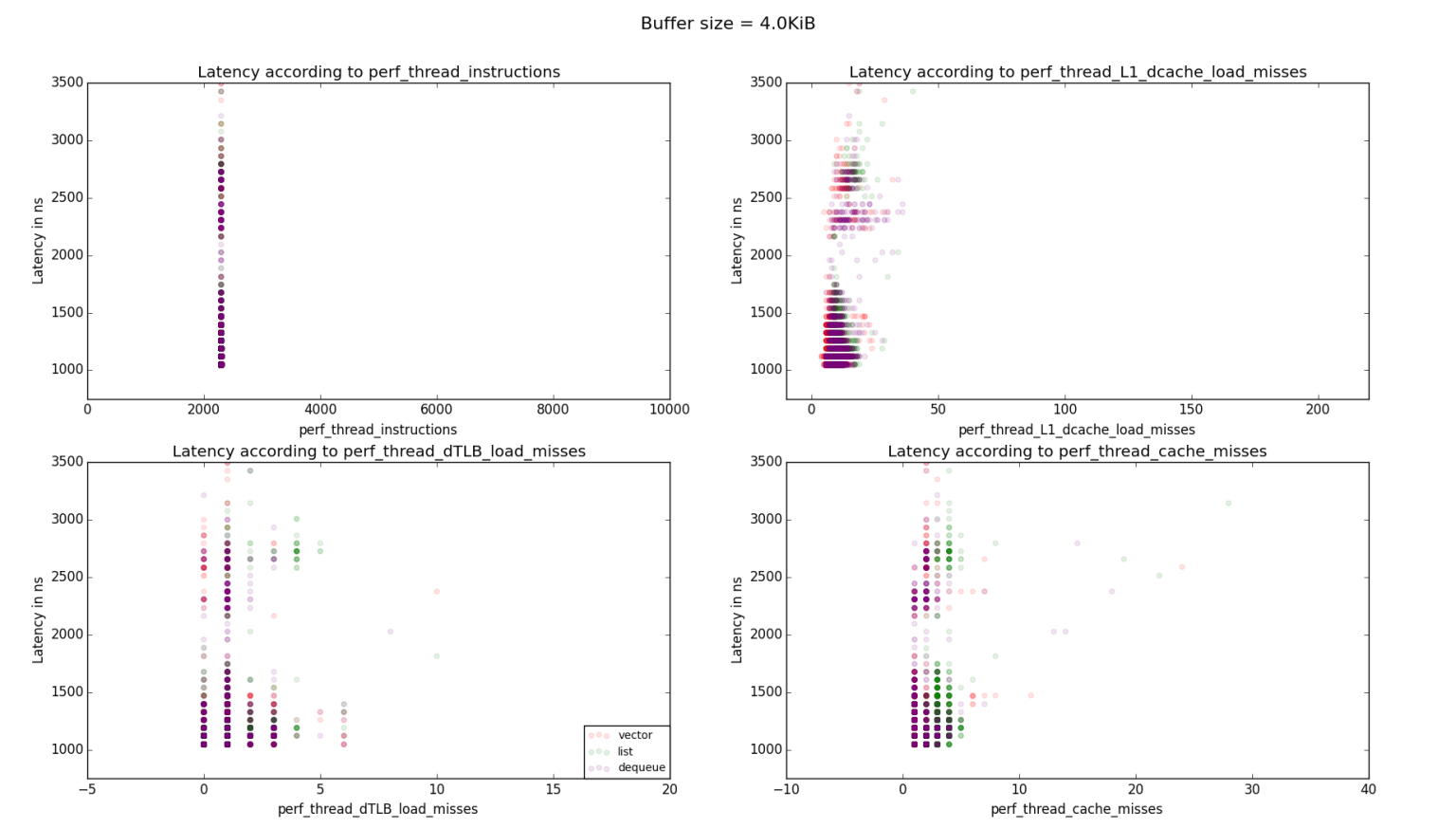
To str() latency



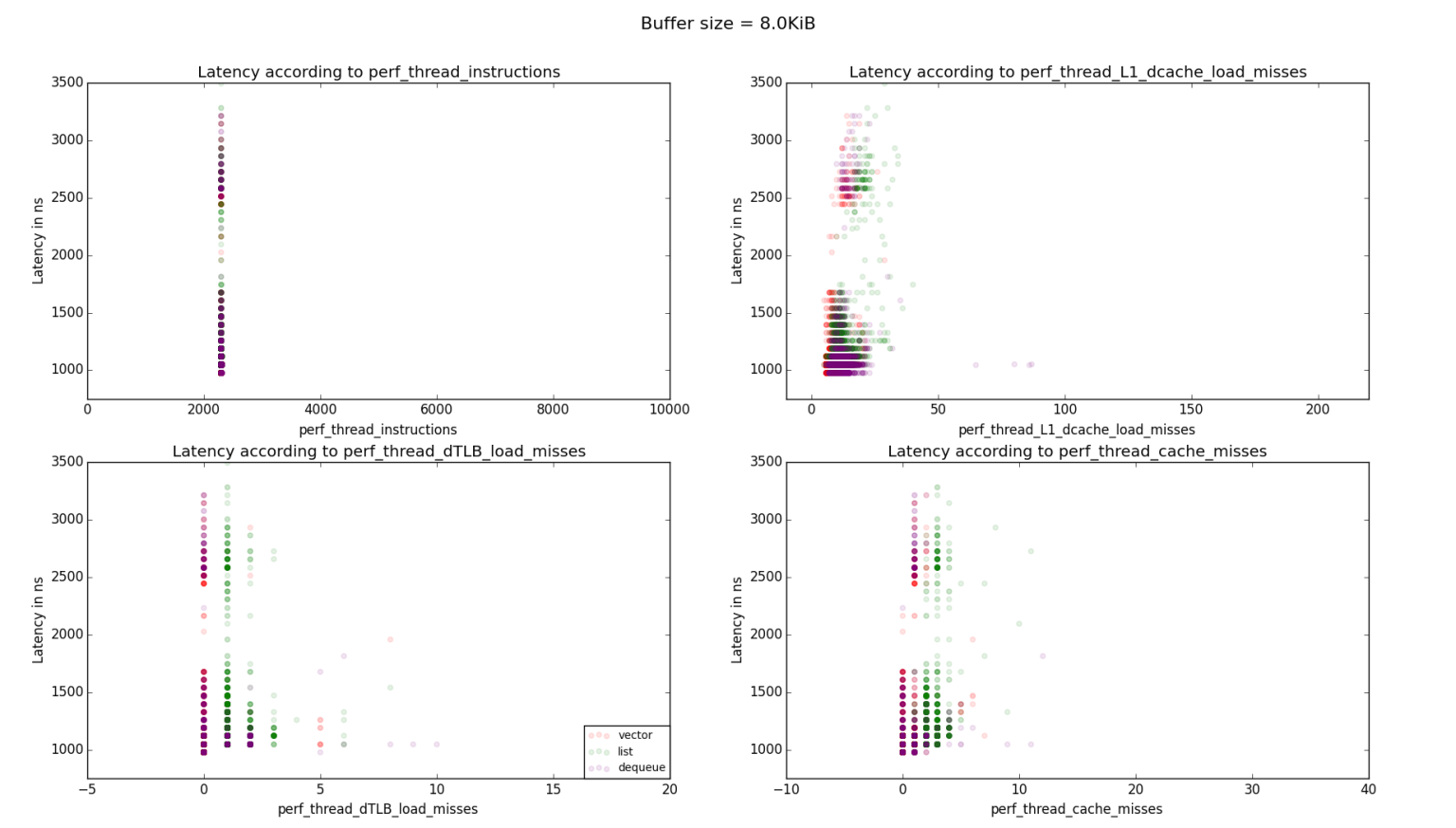
To str() latency



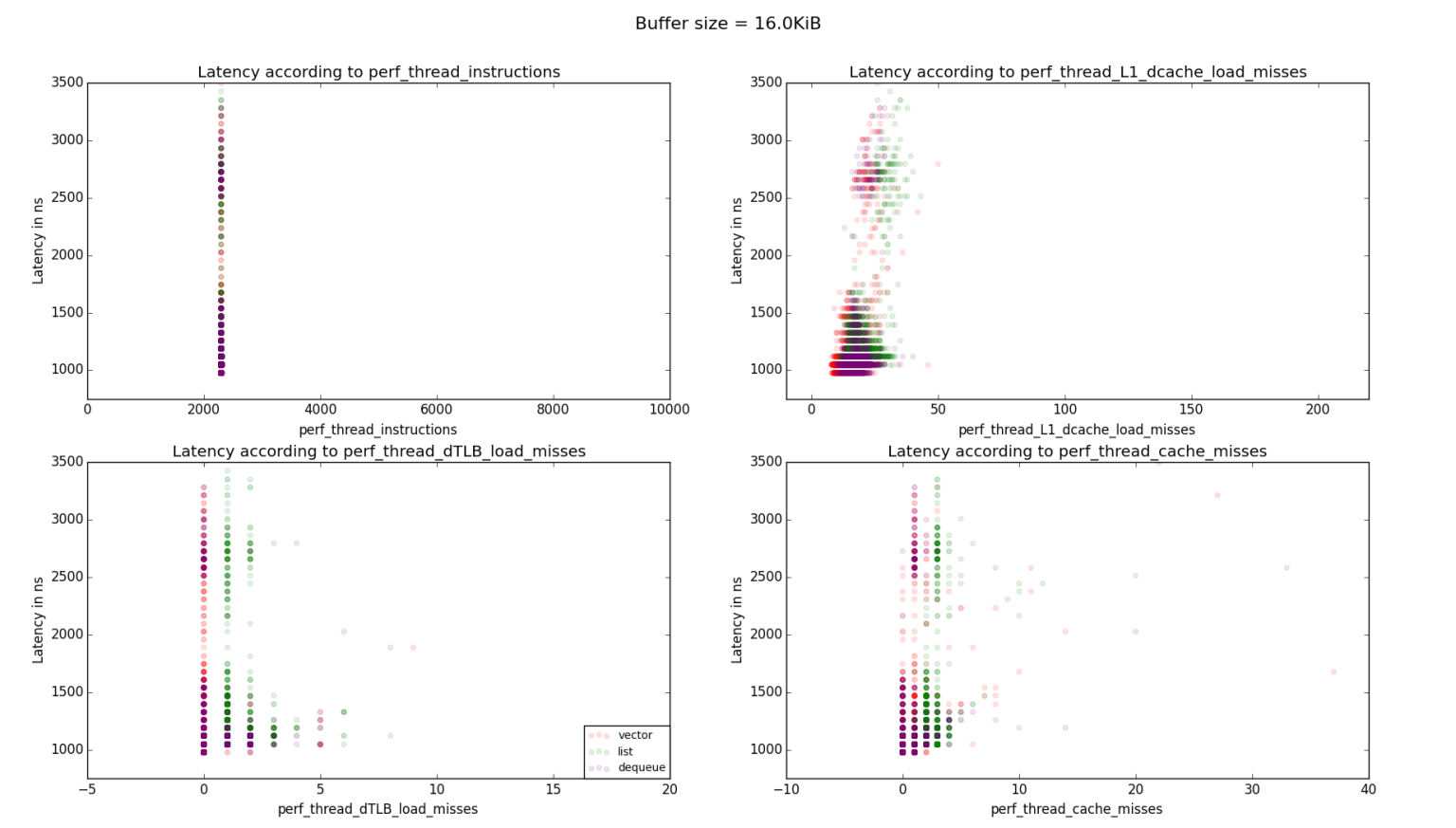
To str() latency



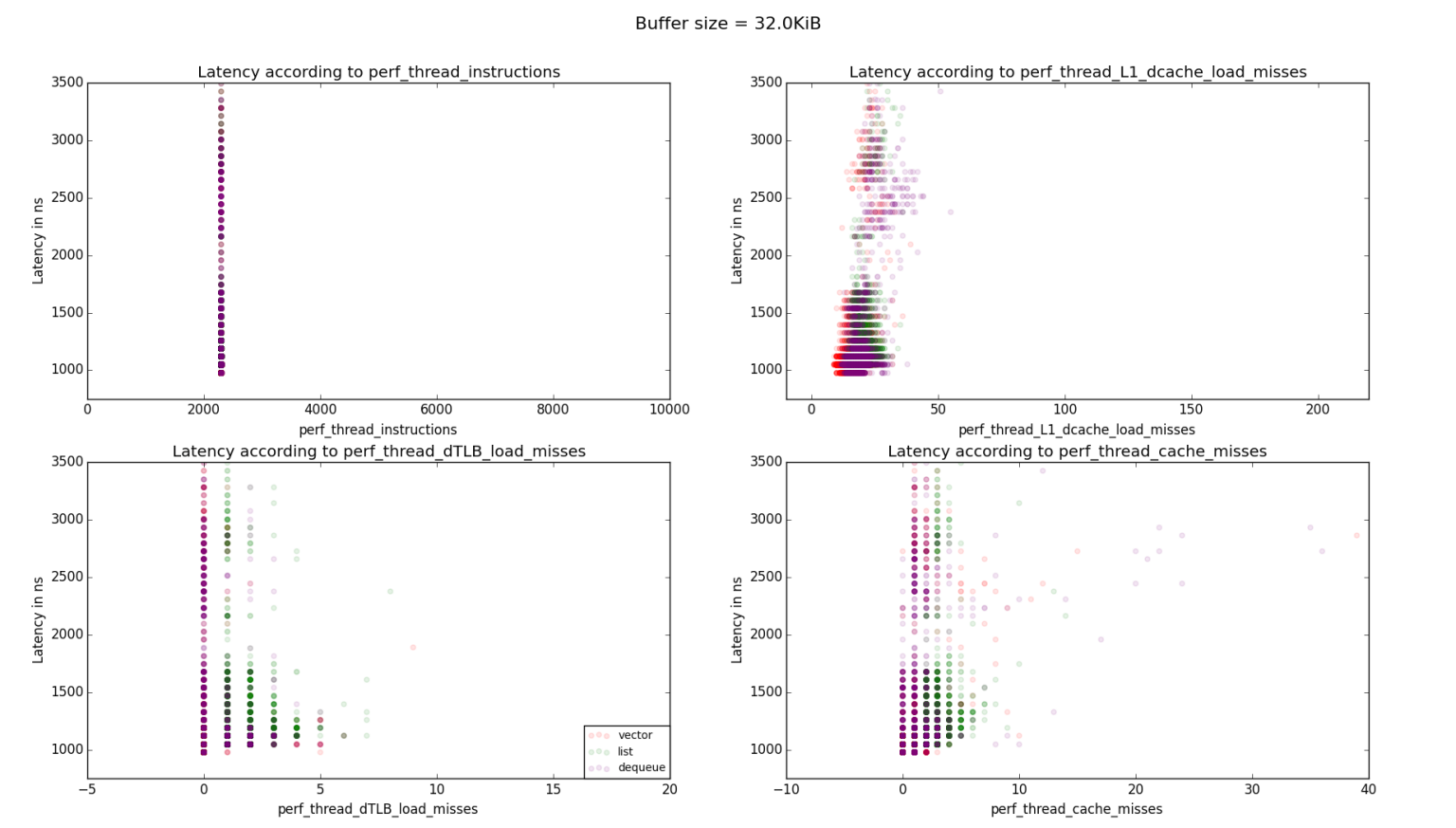
To str() latency



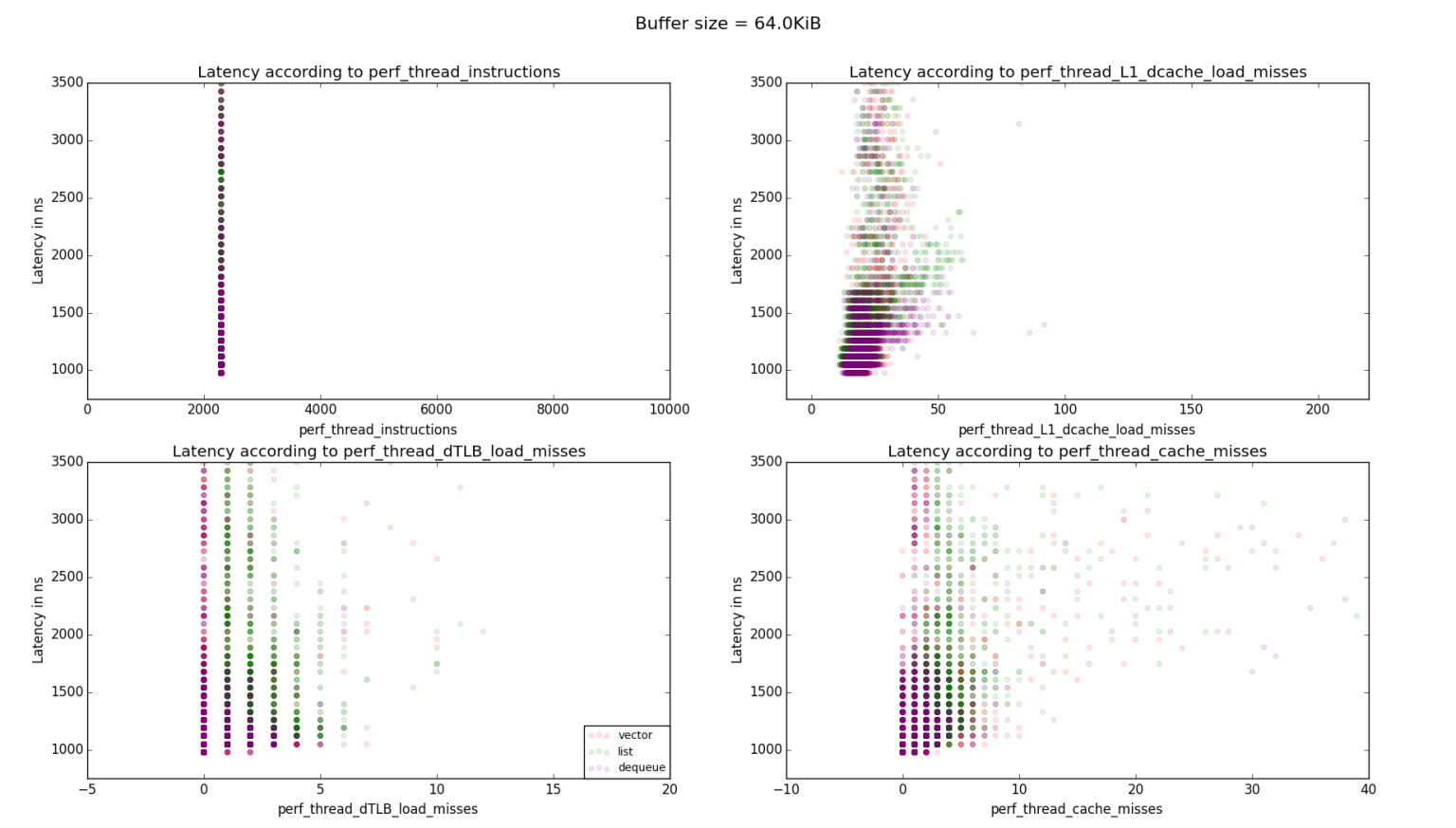
To str() latency



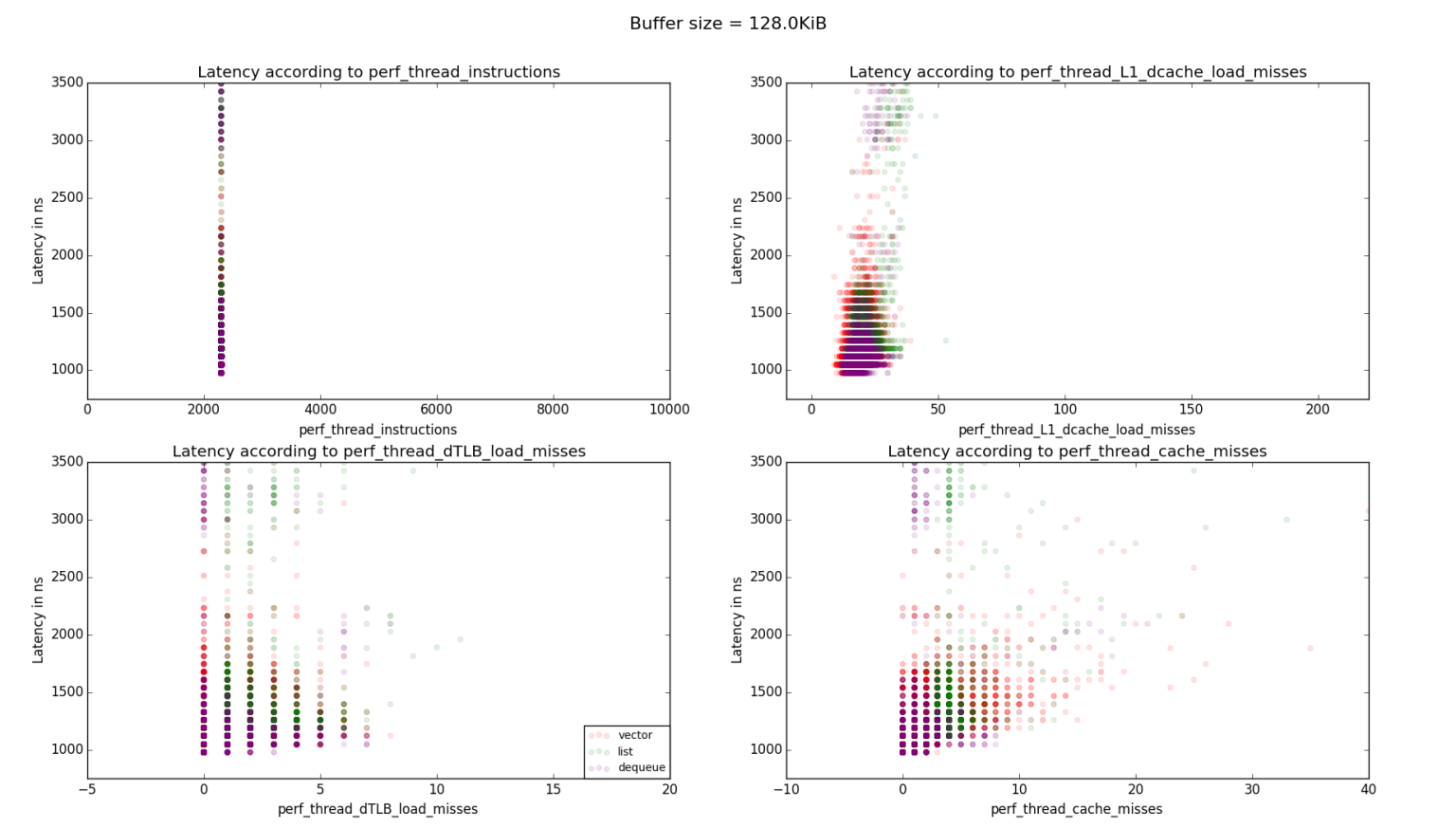
To str() latency



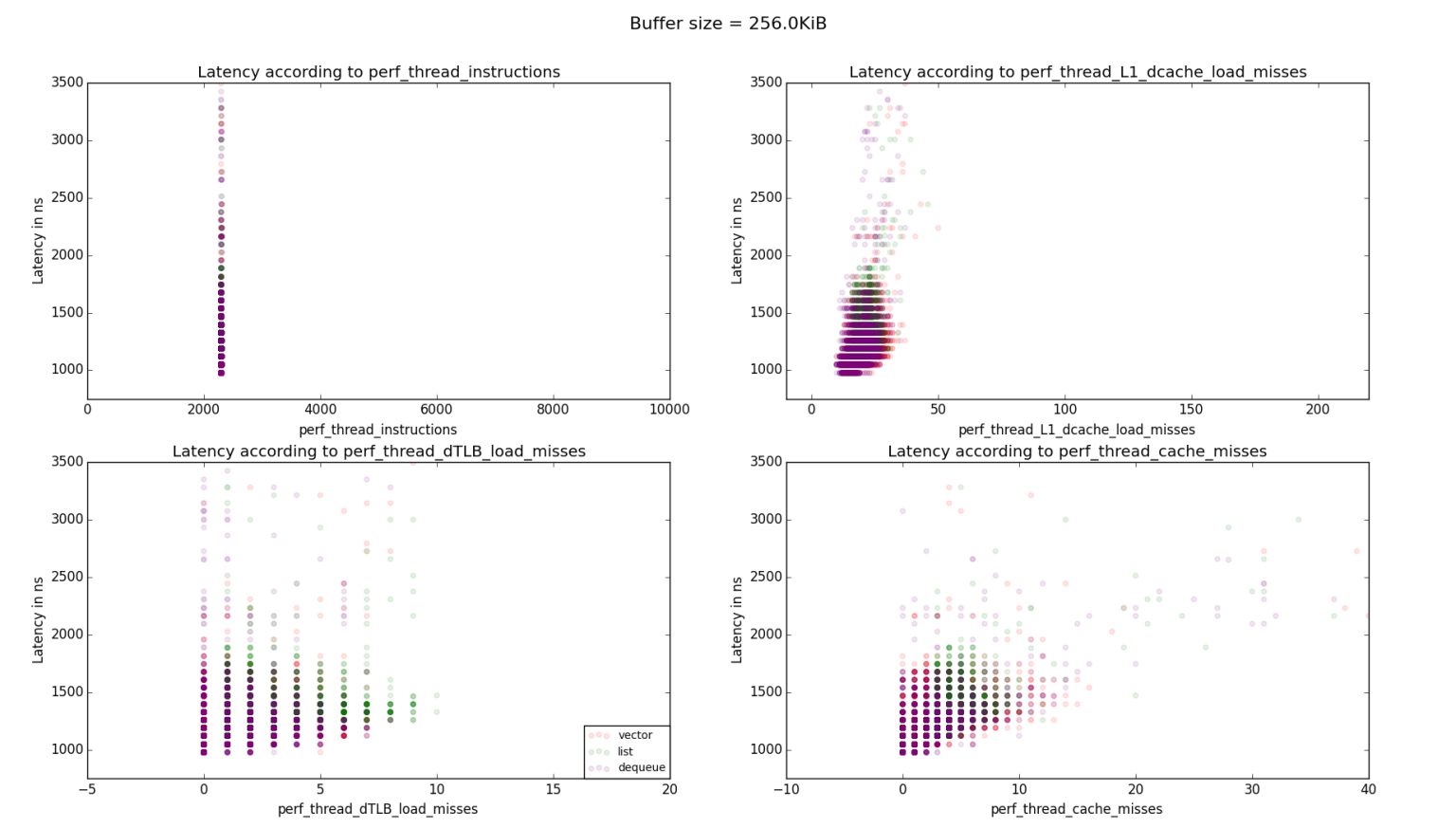
To str() latency



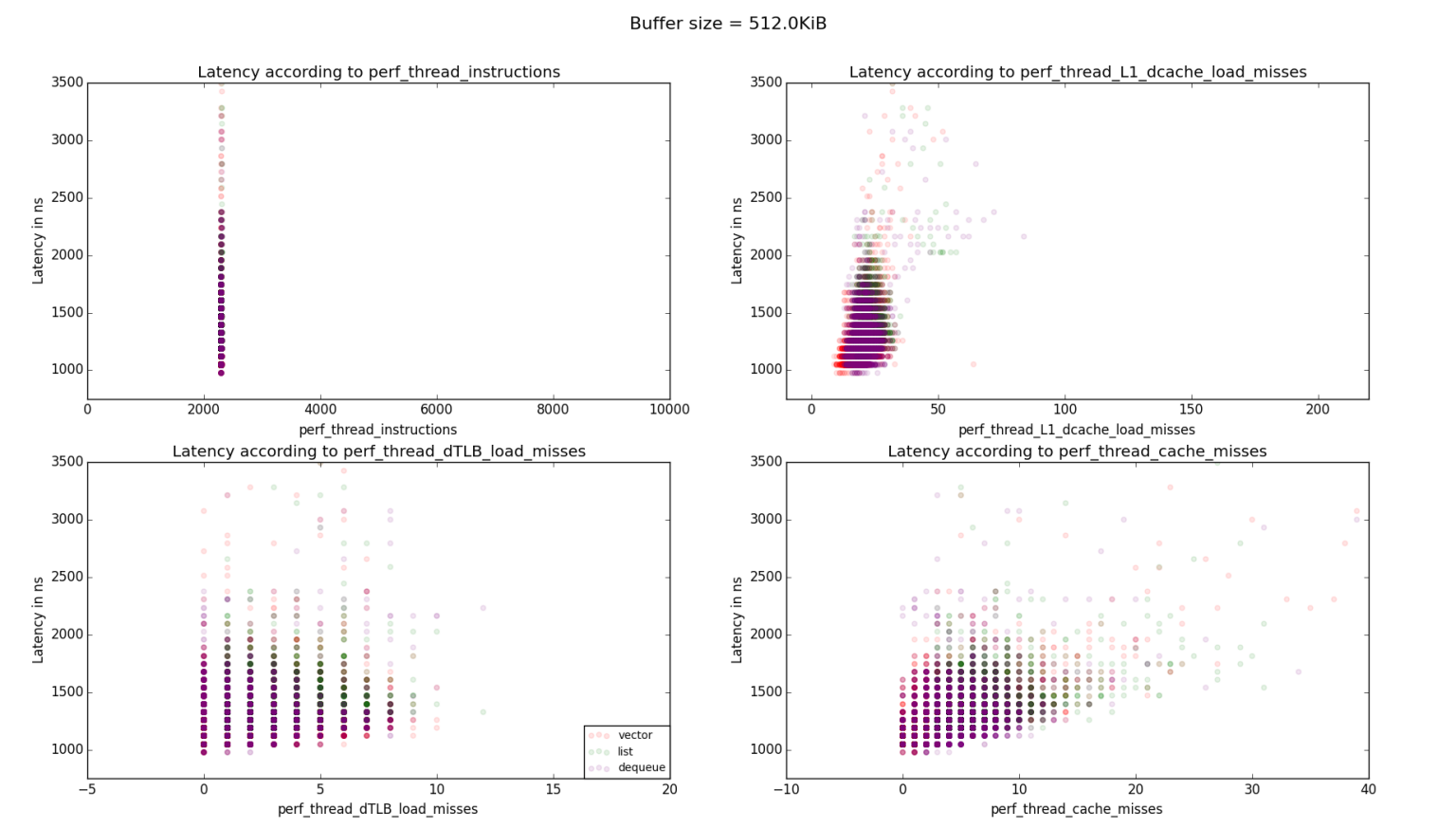
To str() latency



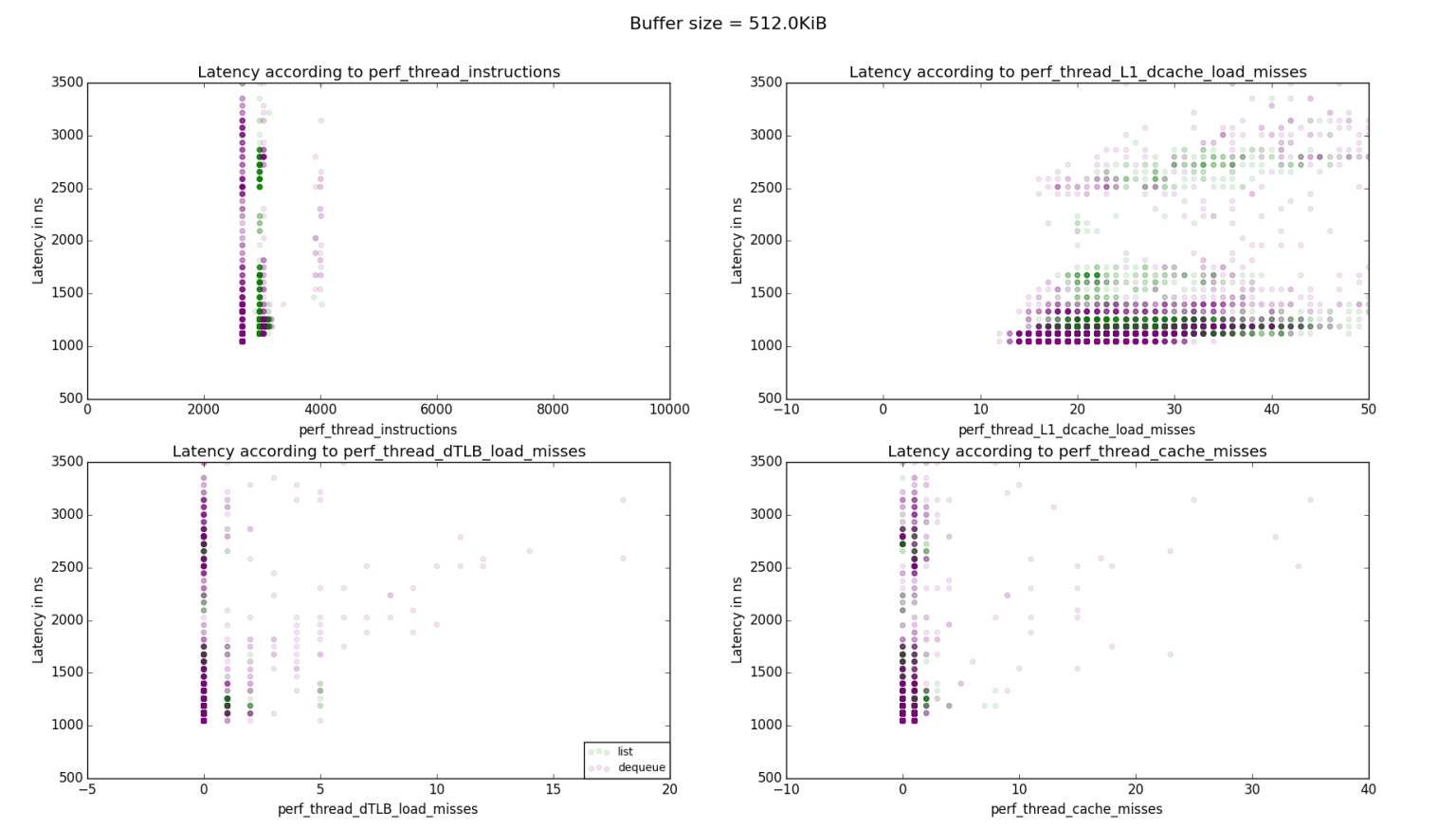
To str() latency



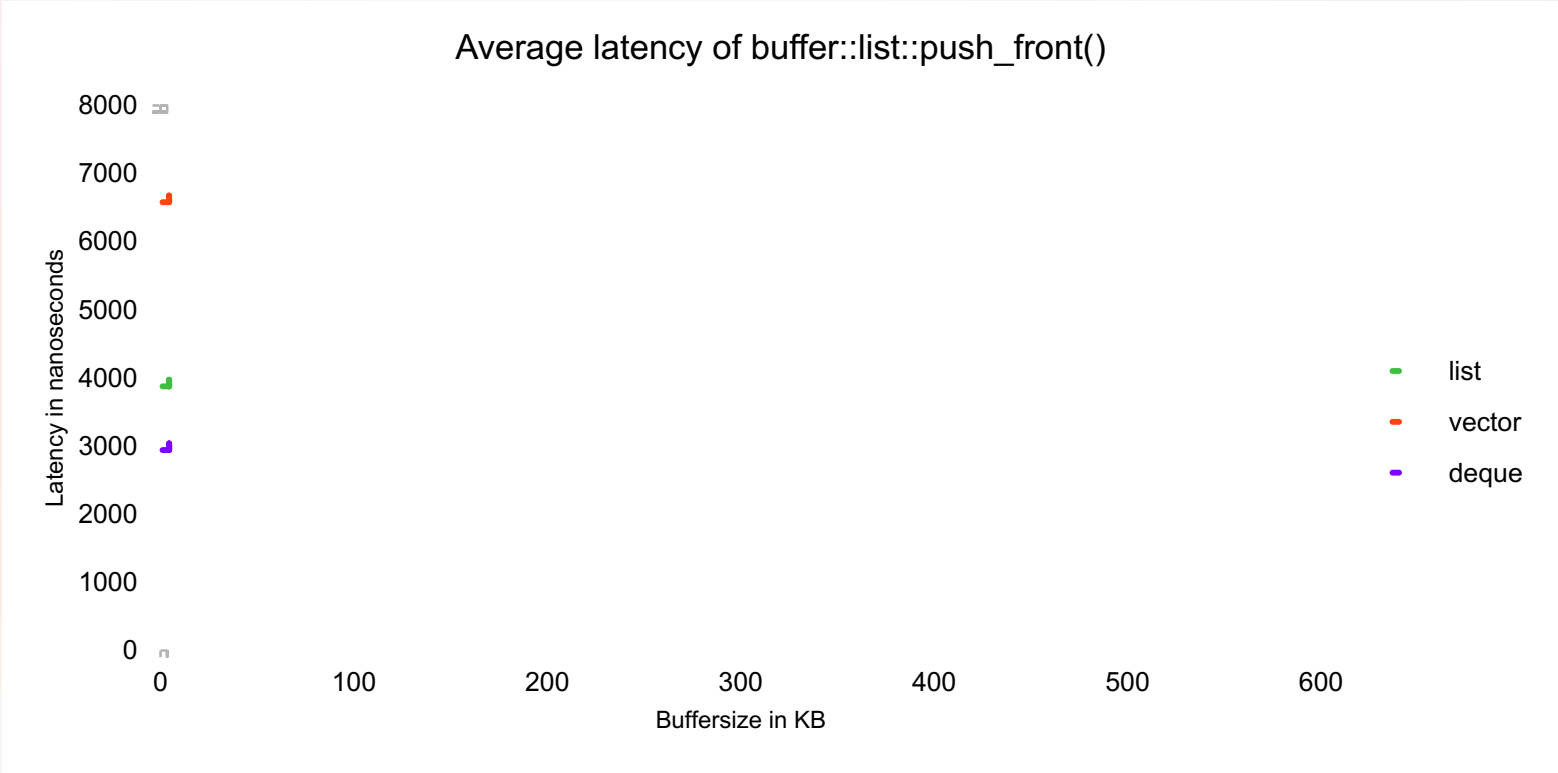
To str() latency



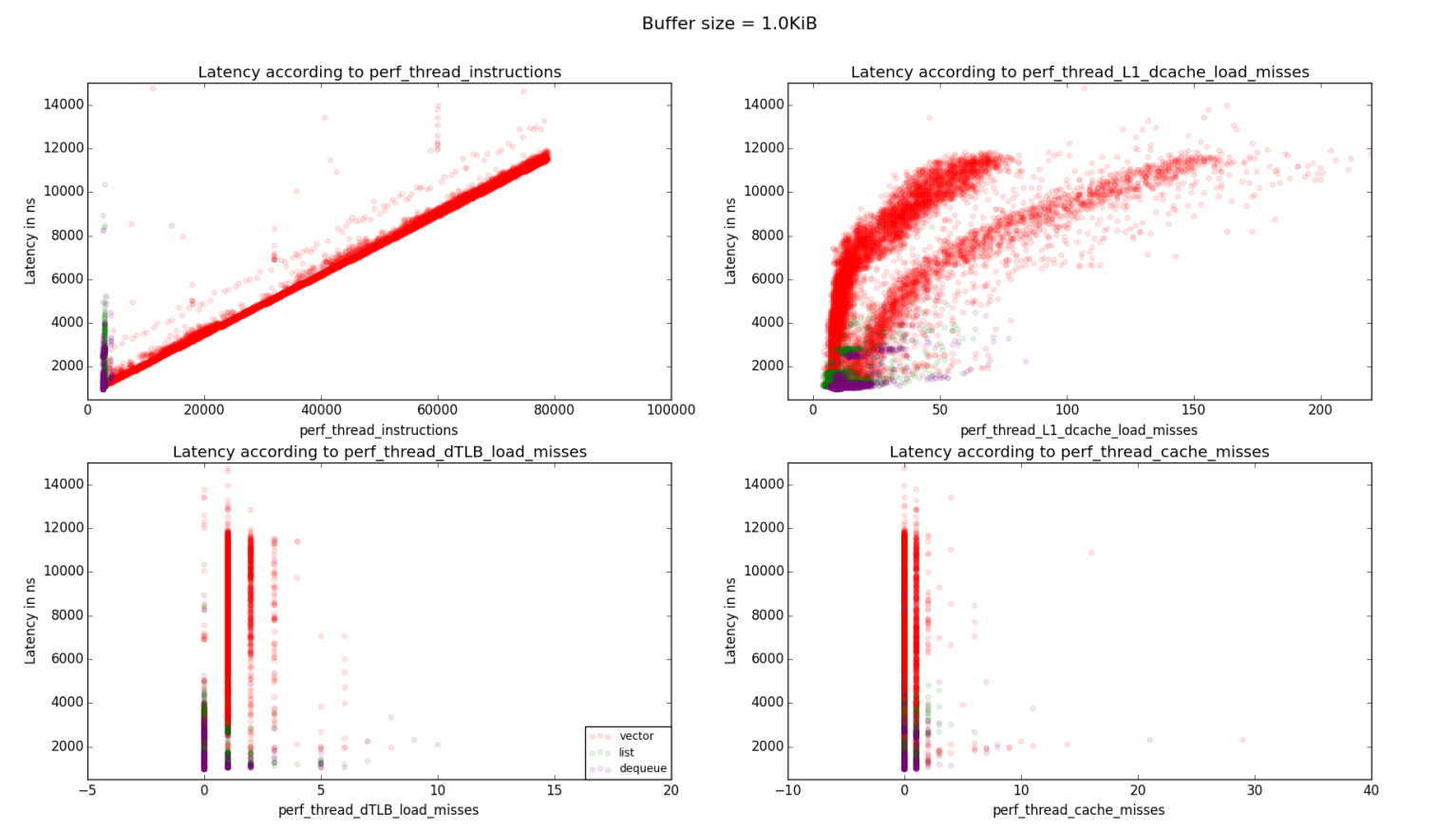
Push front latency



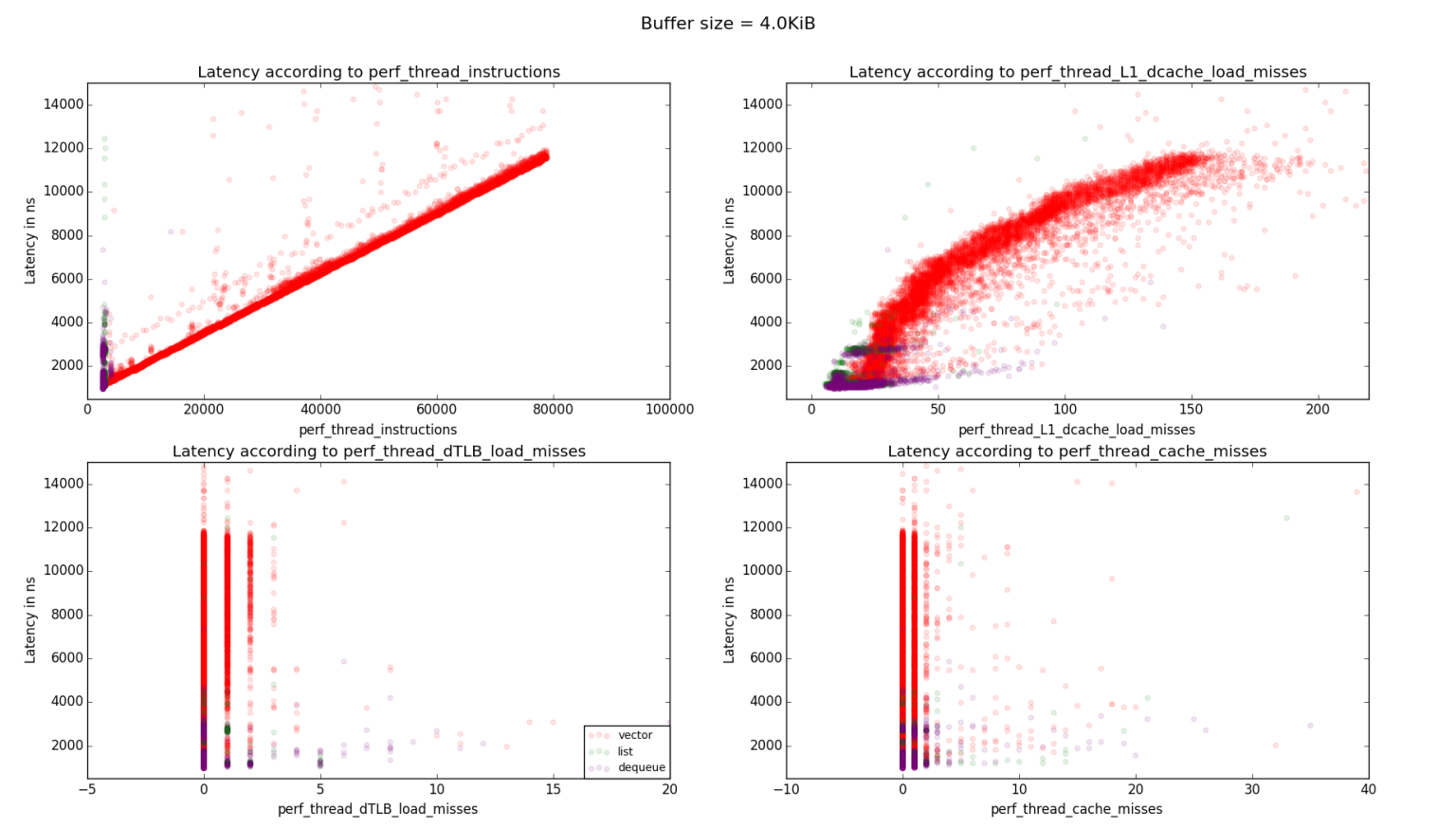
push_front() latency



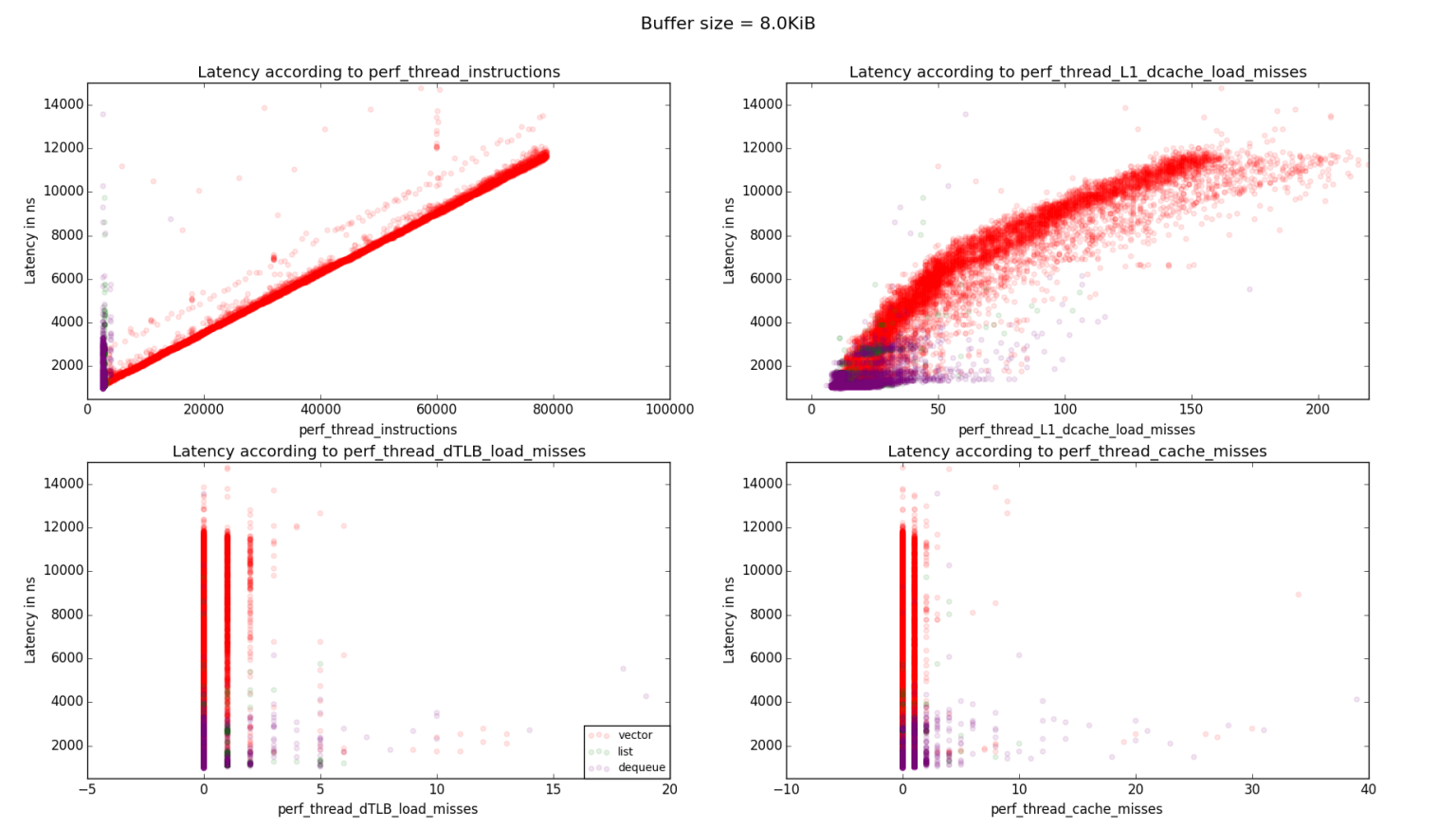
push_front() latency



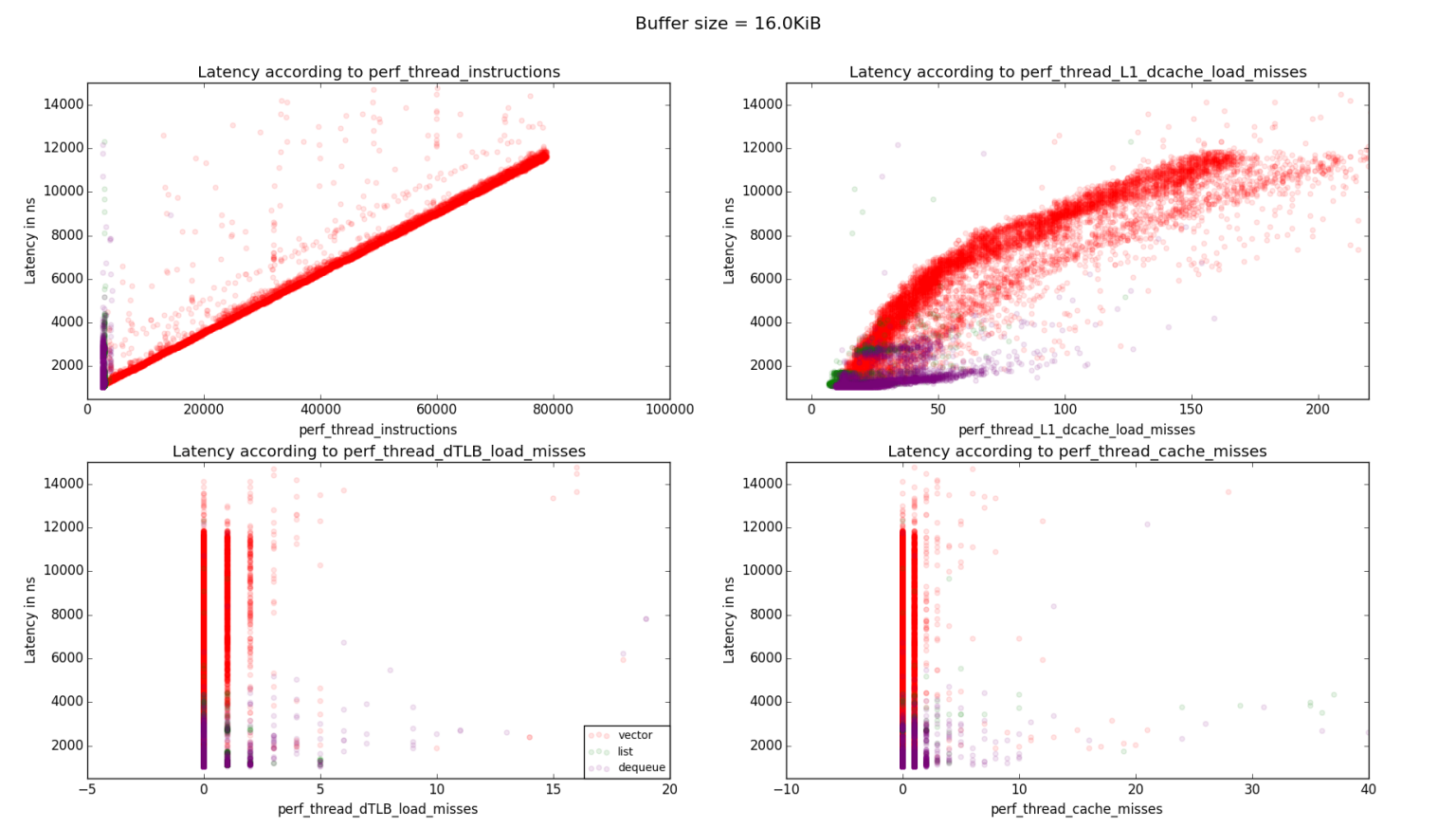
push_front() latency



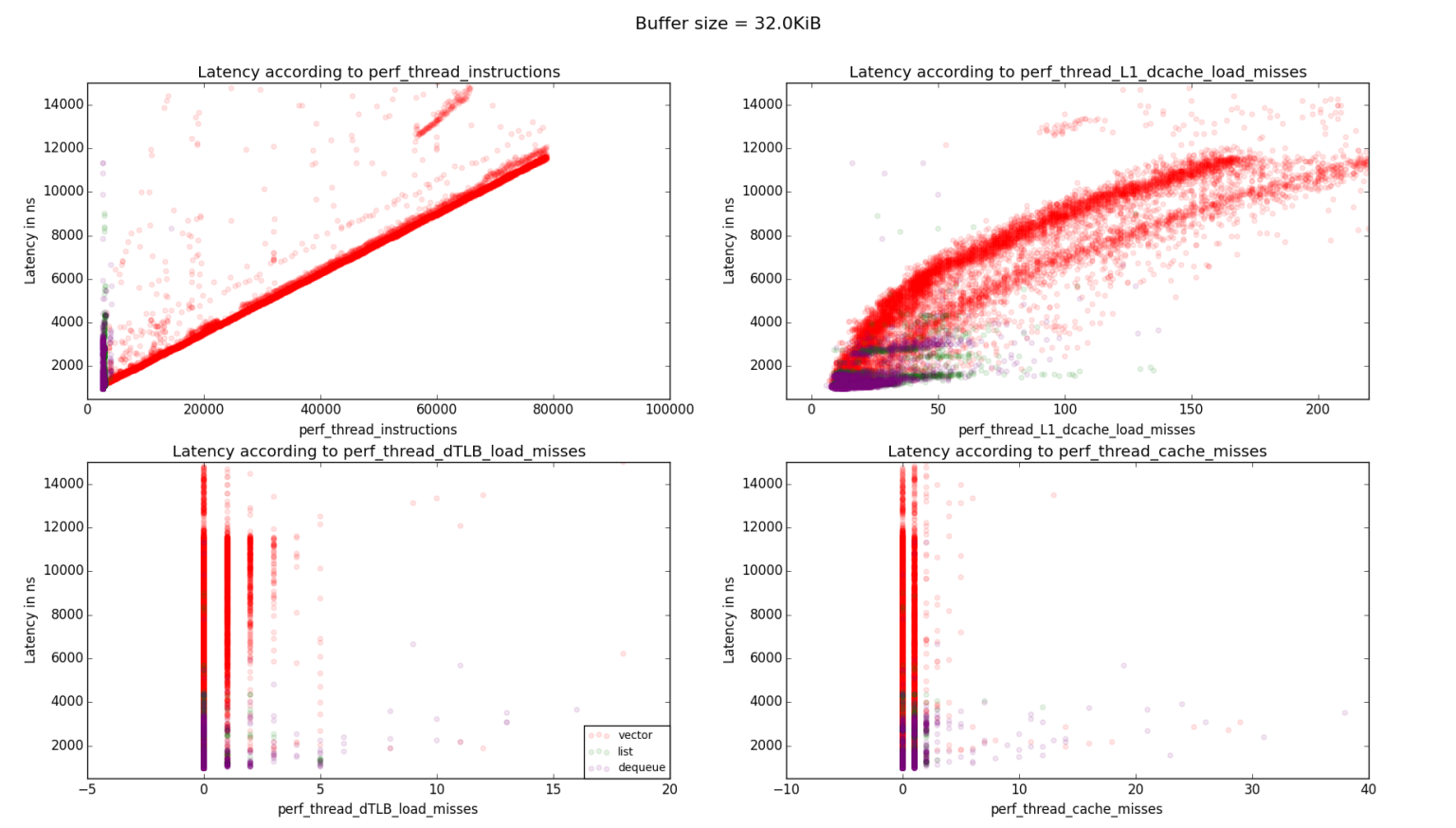
push_front() latency



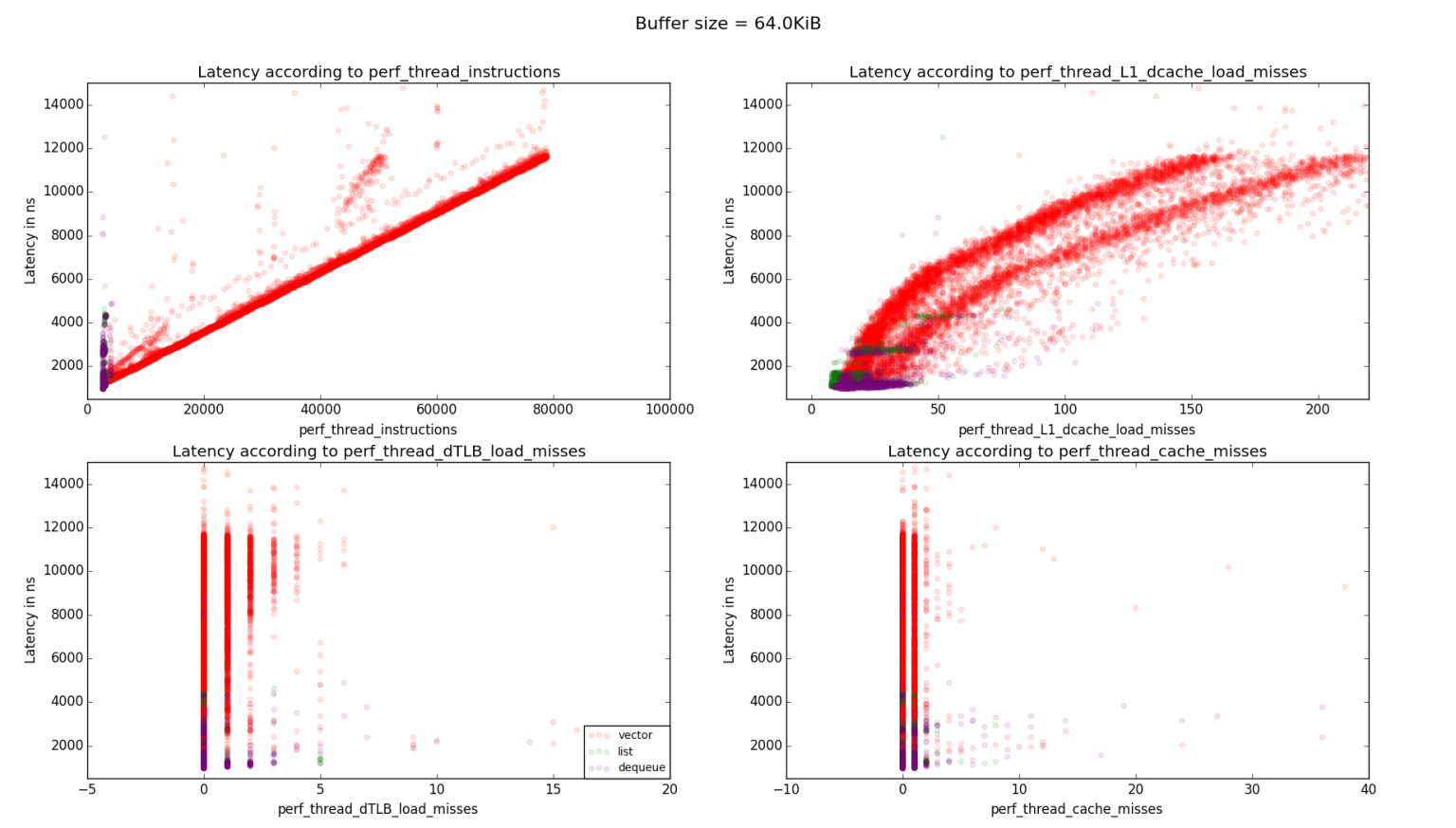
push_front() latency



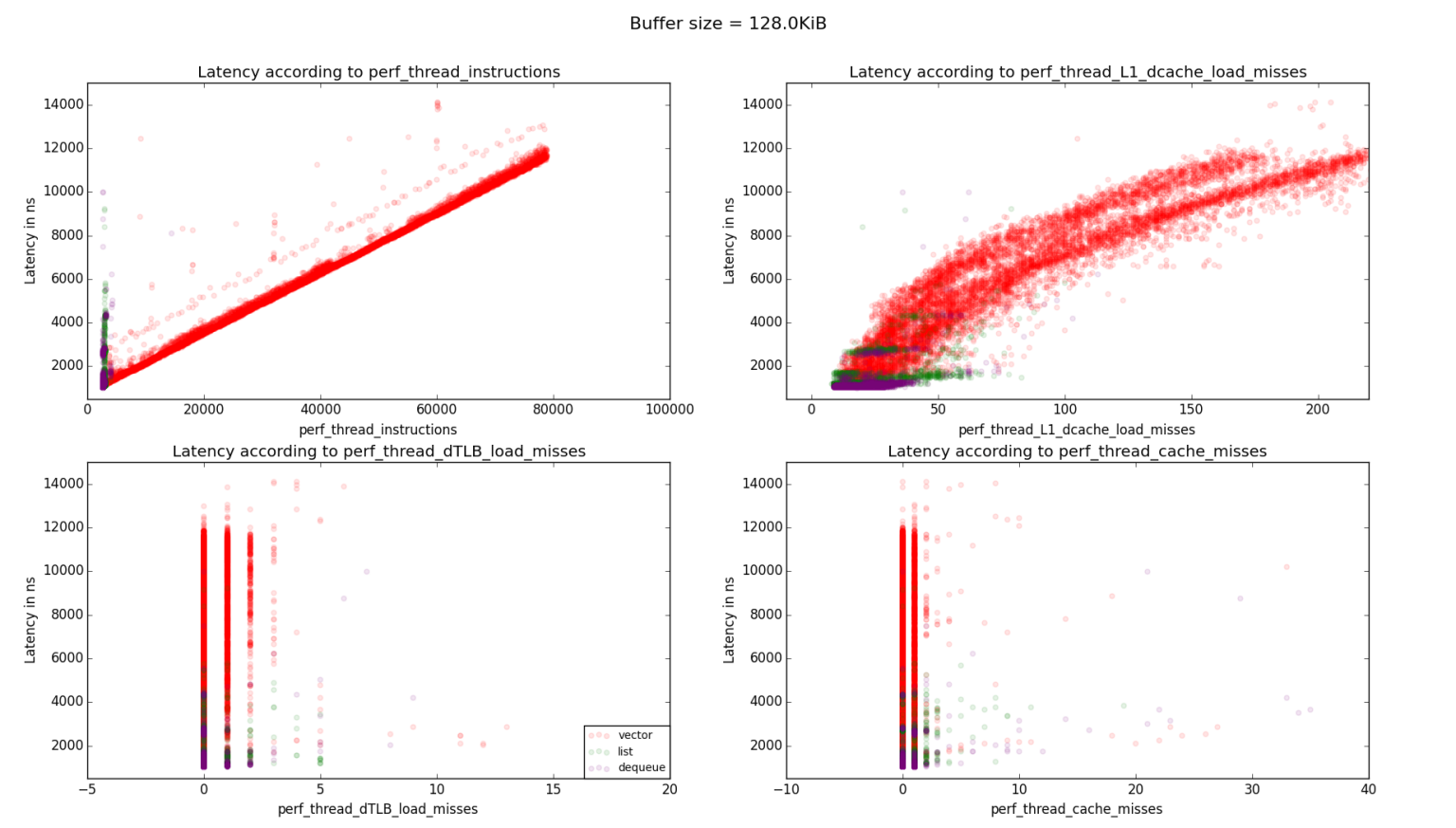
push_front() latency



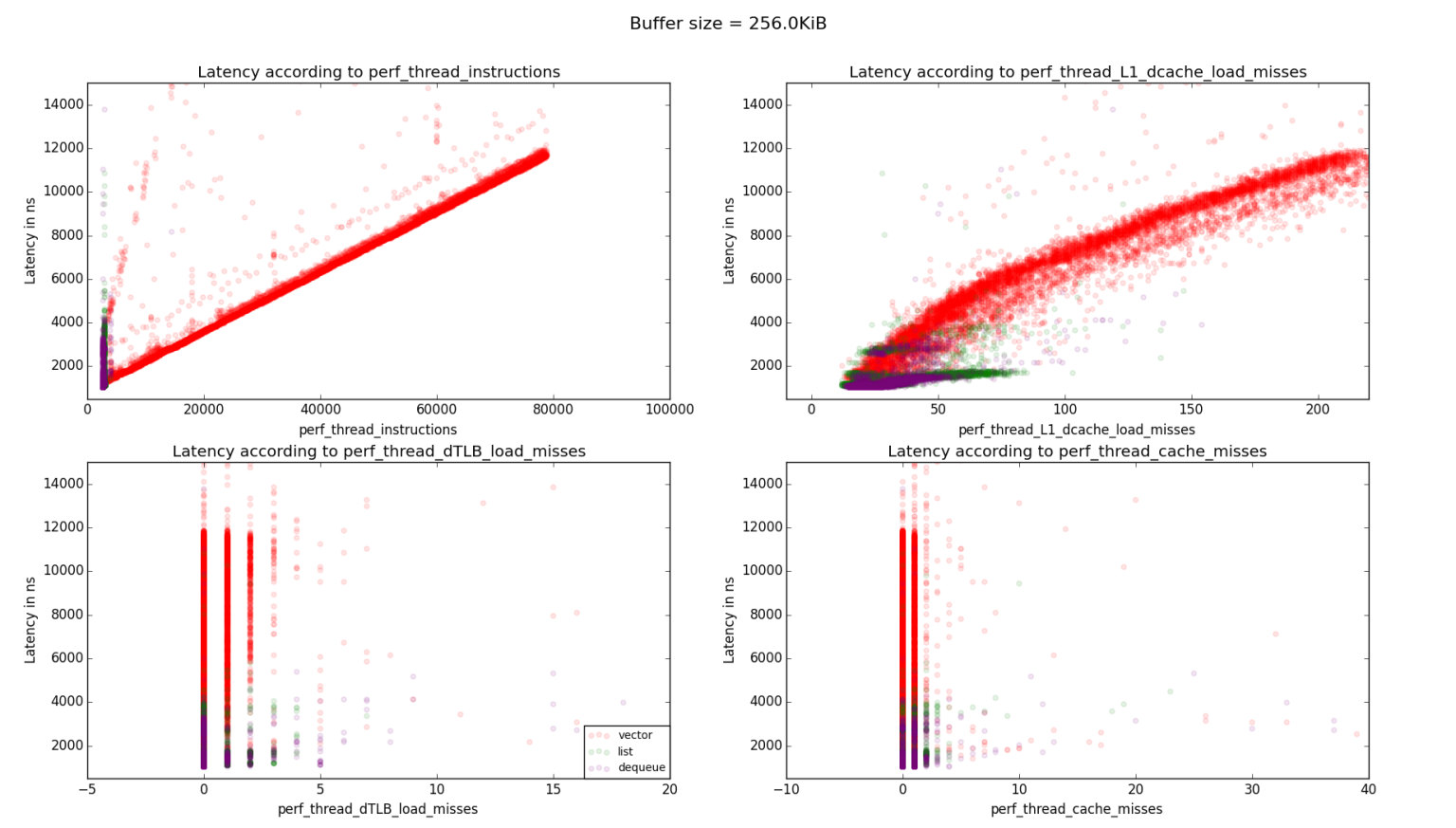
push_front() latency



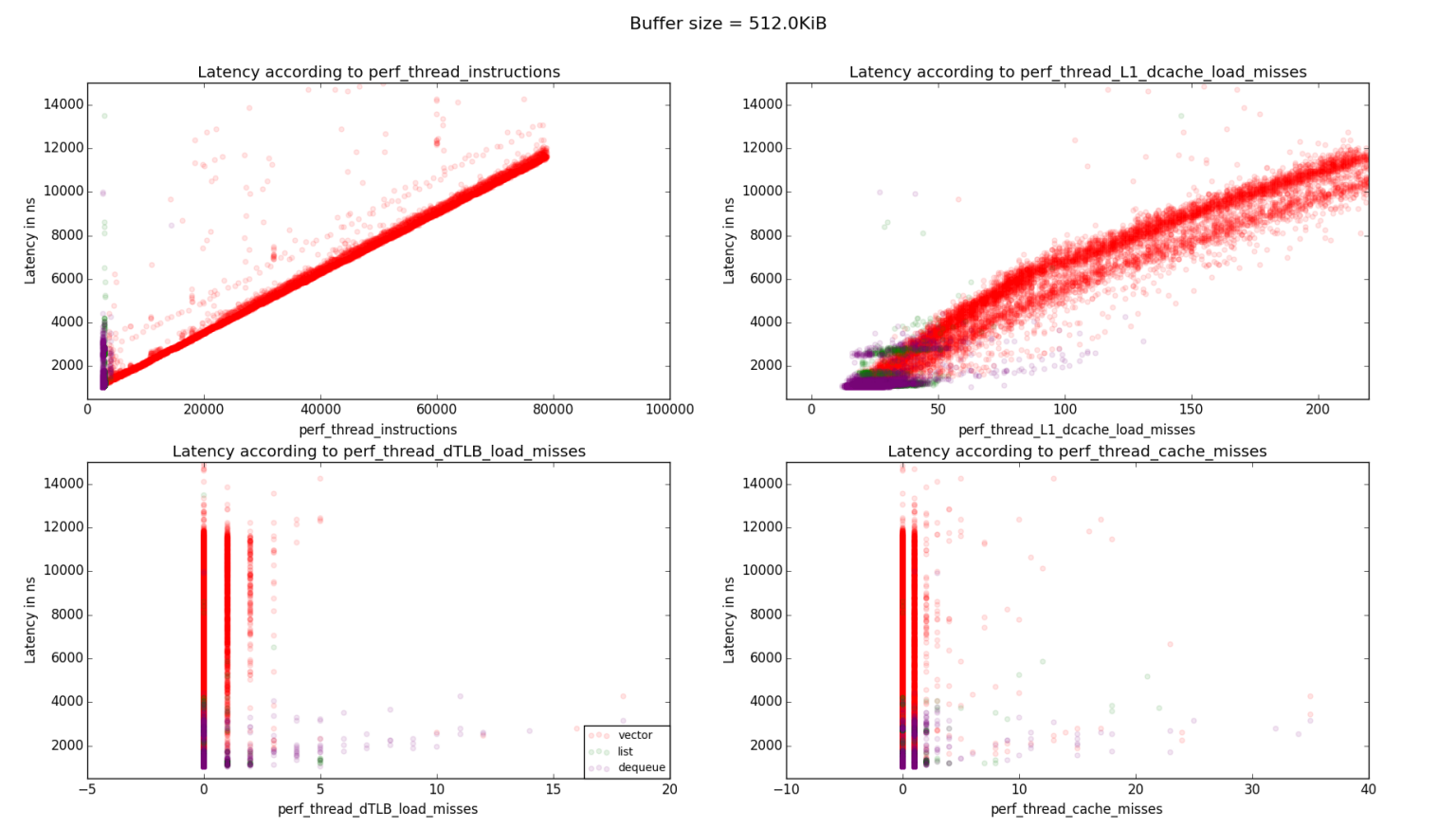
push_front() latency



push_front() latency



push_front() latency



What else?

- LLC misses
- Store cache misses
- Branch prediction misses! (interesting for vector<> and deque<>)
- Page faults
  + minor and major
  + all data structures are similar in a microbenchmark

Trace Compass demo

$> ./do_cmake -DWITH_LTTNG=ON -DWITH_OSD_INSTRUMENT_FUNCTIONS=ON

Or simply

$> ./do_cmake

- The best part: no instrumentation required
- Works with kernel traces from ftrace
- Works with profiling data from perf

Trace Compass demo

Ceph中国社区 | IT大咖说 知识共享平台

DEMO

Thanks to:

- Jesse Williamson (SUSE)
- Adam Emerson (Red Hat)
- The Ceph community

Questions?