



Large Memory Systems, Issues / Fragmentation

Christoph Lameter, Ph.D. R&D Team Lead, Jump Trading LLC
LSF MM Summit, April 23, 2018. Park City, Utah.
cl@linux.com



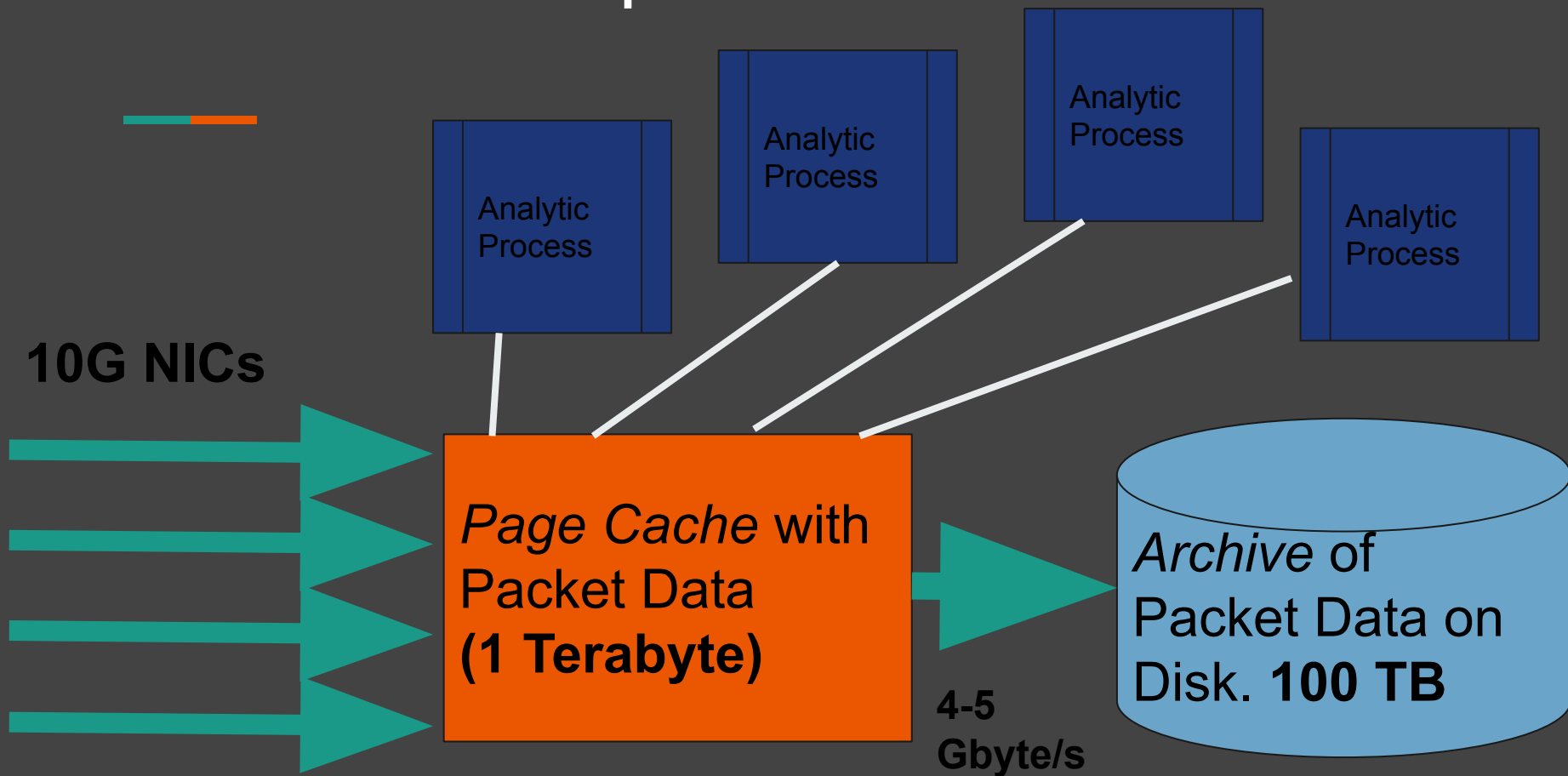


Overview

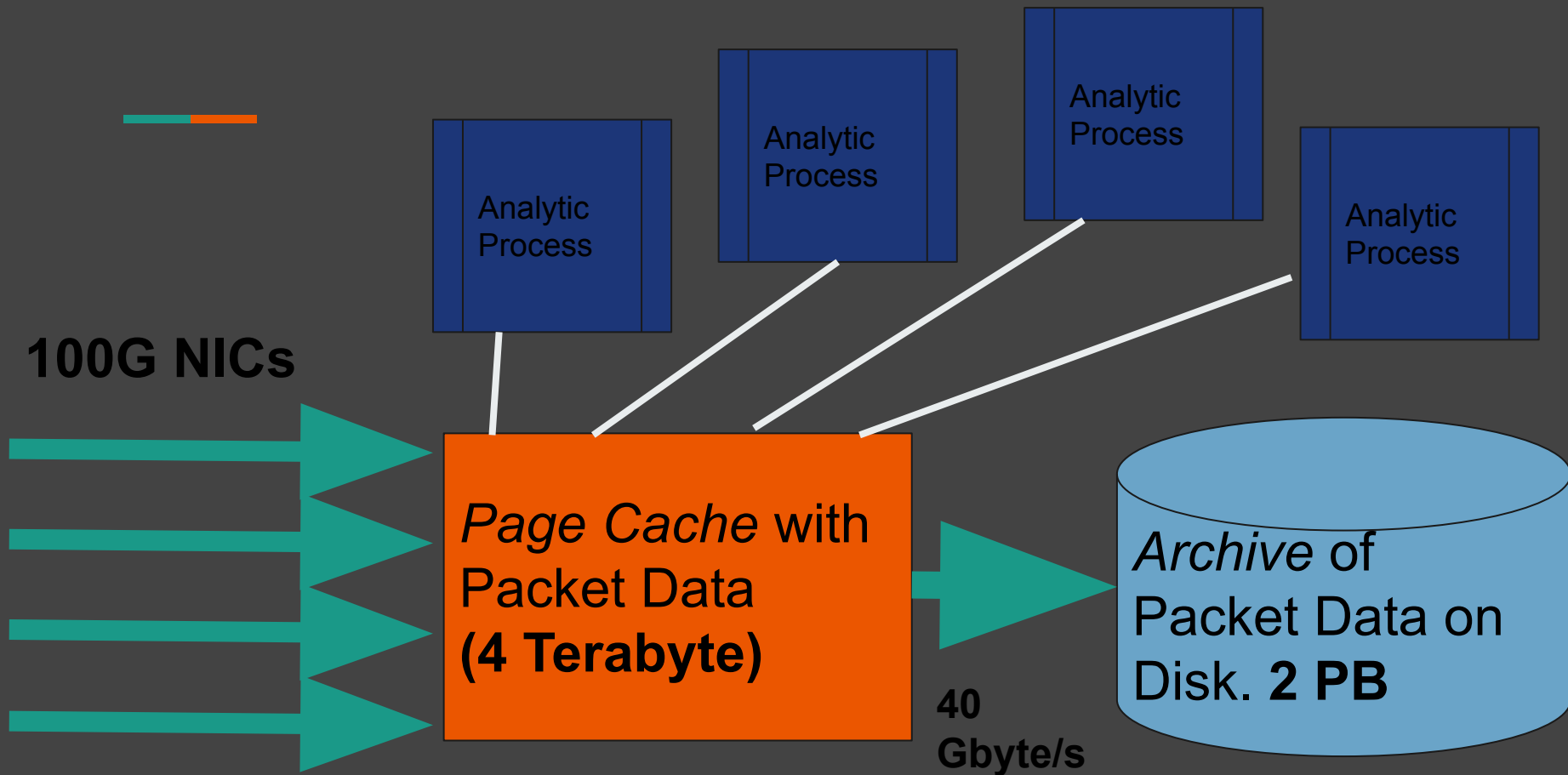
- Large Memory Use case
- The problems
- Solutions and incomplete projects



Current 10 G Packet Capture Solution



Planned 100 G Packet Capture Solution





Some numbers

~ 4Gbyte /sec current max rate to disk from page cache

14.4 TB/hour

10G NIC = 1.25 Gbyte /sec

100G NIC = 12.5 Gbyte/sec

8 hrs of storage ~ 100TB at 10G, 1PB at 100G.





An (Intel) 4K page size issue

SGI Itanium never had that issue.....

Testing on Power 8 also allows > 10GByte/sec to disk from page cache.

Its an overhead per byte issue.

Even a “cp” of a 4Gbyte files shows issues. Now we write custom tools for simple copy operations.

Existing Kernel Techniques:

Faultaround, THP, DAX, ?

Filesystems / Block devices do not support huge pages





Solutions for scaling the page cache

- Increase Base Page Size
 - \ll X size, $N \cdot \text{pte}$ update issues
 - Overallocation and then use segments as needed
- Order N page cache (2007)
 - MMap was never supported
- THP for page cache (2014)
 - No filesystem support?
- Abusing DAX to create huge mappings (via memmap= parameter)
- Using Nonvolatile memory and thus not abusing DAX.
- CMA(???) on intel to make memblocks movable?



Key MM issue: Contiguous physical Memory

- 01 | Reservation scheme (as for huge pages)
- 02 | Solve the fragmentation issue.
- 03 | Boot time reservations
- 04 | Virtually mapping via page tables
- 05 | Garbage Collection

Problems to consider:

Boot time reservations for larger segments


Reboot to get contiguous memory

System degrades when it is running

4K wastes of memory due to page structs.

4K Page structs cause increased cache footprint and processing overhead.

4TB memory requires 1 billion page structs. DONT!



Enhancements to support contiguous memory

- Enable reliable allocation of large contiguous memory
 - Reclaim any object
 - Movable objects in particular inodes and dentries
- Sample xarray cache with movable objects through slab allocator enhancements
- Reservation scheme for diverse sizes of contiguous memory segments.
- MAP_CONTIG support in mmap
- Page allocator statistics on allocations and failures for various contiguous memory sizes.
- Garbage collection like in Java. Mark all pointers.
-



Thank you.

