Efficient parallel programming on ccNUMA nodes

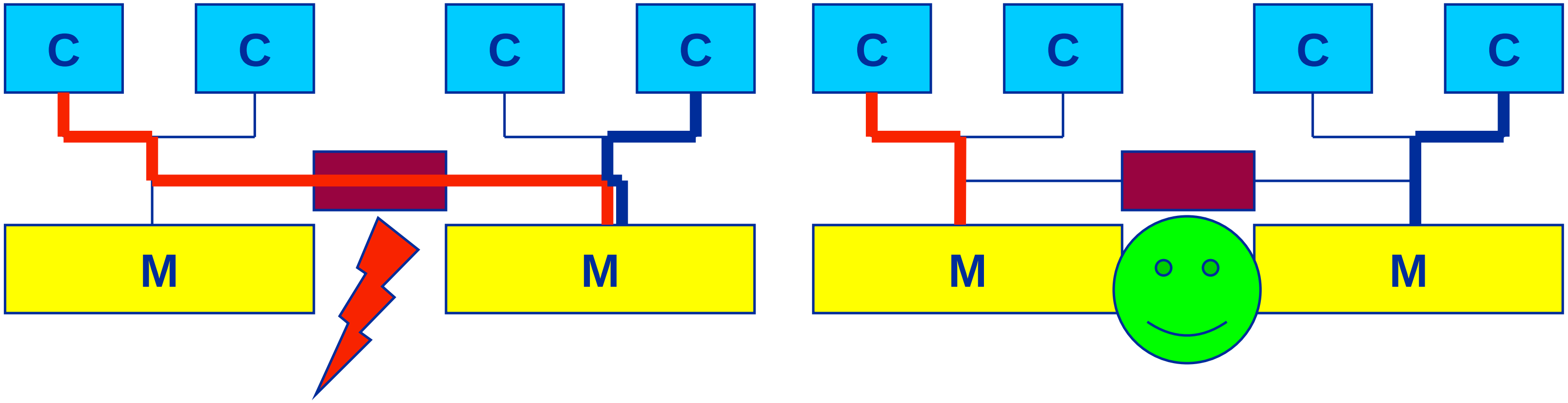
Performance characteristics of ccNUMA nodes

First touch placement policy

ccNUMA performance problems

"The other affinity" to care about

- **ccNUMA:**
  - Whole memory is transparently accessible by all processors
  - but physically distributed
  - with varying bandwidth and latency
  - and potential contention (shared memory paths)
- **How do we make sure that memory access is always as "local" and "distributed" as possible?**

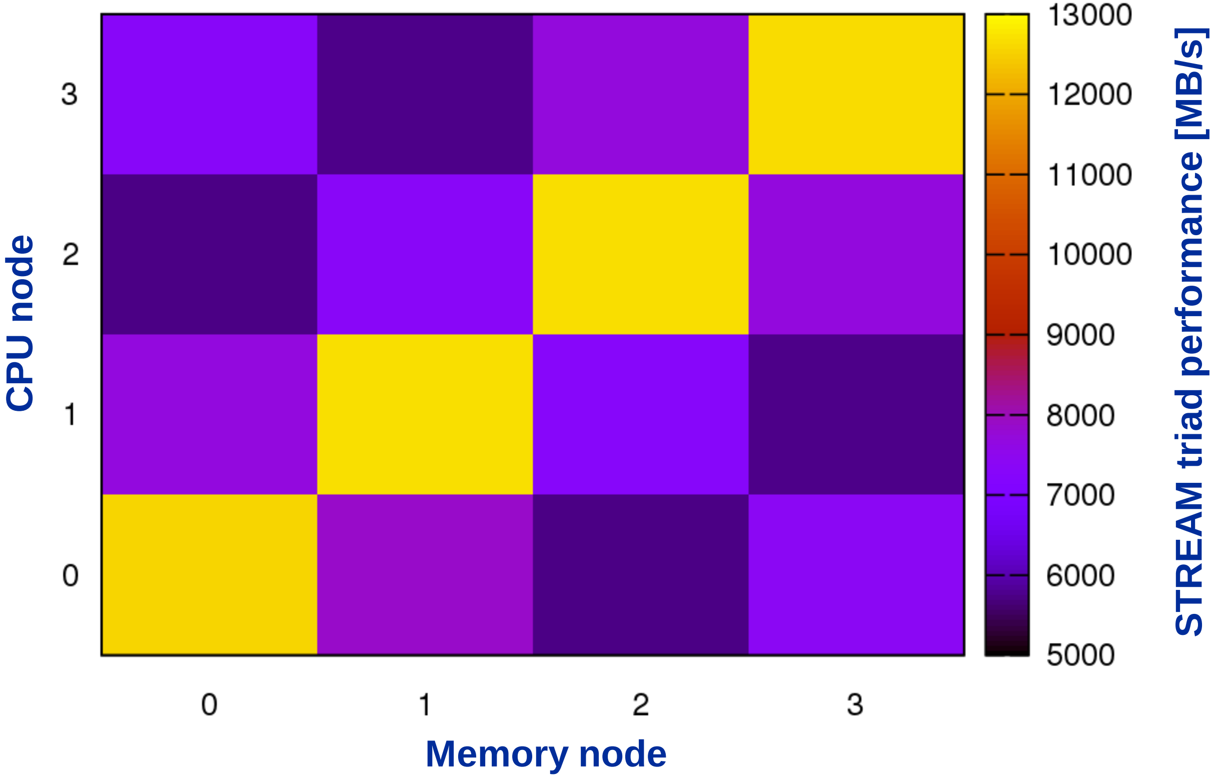


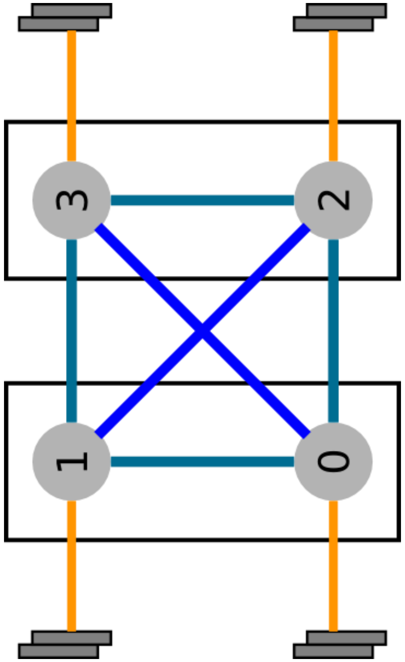
- Page placement is implemented in units of OS pages (often 4kB, possibly more)

Cray XE6 Interlagos node

4 chips, two sockets, 8 threads per ccNUMA domain

- **ccNUMA map: Bandwidth penalties for remote access**
  - Run 8 threads per ccNUMA domain (1 chip)
  - Place memory in different domain → 4x4 combinations
  - STREAM triad benchmark using nontemporal stores





numactl as a simple ccNUMA locality tool :
How do we enforce some locality of access?

- `numactl` can influence the way a binary maps its memory pages:

```
numactl --membind=<nodes> a.out     # map pages only on <nodes>
        --preferred=<node>  a.out   # map pages on <node>
                                    # and others if <node> is full
        --interleave=<nodes> a.out  # map pages round robin across
                                    # all <nodes>
```

- Examples:

ccNUMA map scan

```
for m in `seq 0 3`; do
  for c in `seq 0 3`; do
    env OMP_NUM_THREADS=8 \
        numactl --membind=$m --cpunodebind=$c ./stream
  enddo
enddo
```

```
env OMP_NUM_THREADS=4 numactl --interleave=0-3 \
                           likwid-pin -c N:0,4,8,12 ./stream
```

- But what is the default without `numactl?`

ccNUMA default memory locality

- **"Golden Rule" of ccNUMA:**

  A memory page gets mapped into the local memory of the processor that first touches it!

  - Except if there is not enough local memory available
  - This might be a problem, see later
- **Caveat: "touch" means "write", not "allocate"**
- **Example:**

```
double *huge = (double*)malloc(N*sizeof(double));

for(i=0; i<N; i++) // or i+=PAGE_SIZE/sizeof(double)
    huge[i] = 0.0;
```

Memory not mapped here yet

Mapping takes place here

- **It is sufficient to touch a single item to map the entire page**

Coding for ccNUMA data locality

- Most simple case: explicit initialization

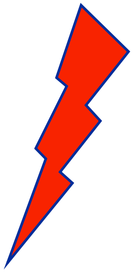
```
integer,parameter :: N=10000000
double precision A(N), B(N)



A=0.d0



!$OMP parallel do
do i = 1, N
  B(i) = function ( A(i) )
end do
!$OMP end parallel do
```



```
integer,parameter :: N=10000000
double precision A(N),B(N)
!$OMP parallel
!$OMP do schedule(static)
do i = 1, N
  A(i)=0.d0
end do
!$OMP end do
...
!$OMP do schedule(static)
do i = 1, N
  B(i) = function ( A(i) )
end do
!$OMP end do
!$OMP end parallel
```

Coding for ccNUMA data locality

- Sometimes initialization is not so obvious: I/O cannot be easily parallelized, so “localize” arrays before I/O

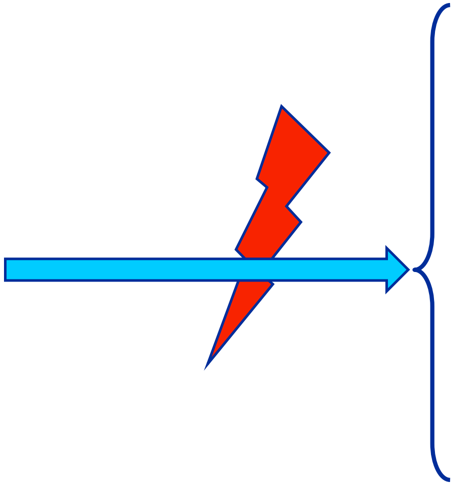
```
integer,parameter :: N=10000000
double precision A(N), B(N)




READ(1000) A




!$OMP parallel do
do i = 1, N
  B(i) = function ( A(i) )
end do
!$OMP end parallel do
```



```
integer,parameter :: N=10000000
double precision A(N),B(N)
!$OMP parallel
!$OMP do schedule(static)
do i = 1, N
  A(i)=0.d0
end do
!$OMP end do
!$OMP single
READ(1000) A
!$OMP end single
!$OMP do schedule(static)
do i = 1, N
  B(i) = function ( A(i) )
end do
!$OMP end do
!$OMP end parallel
```

Coding for Data Locality

- **Required condition: OpenMP loop schedule of initialization must be the same as in all computational loops**
  - Only choice: `static`! Specify explicitly on all NUMA-sensitive loops, just to be sure…
  - Imposes some constraints on possible optimizations (e.g. load balancing)
  - Presupposes that all worksharing loops with the same loop length have the same thread-chunk mapping
  - If dynamic scheduling/tasking is unavoidable, more advanced methods may be in order
    - See below
- **How about global objects?**
  - Better not use them
  - If communication vs. computation is favorable, might consider properly placed copies of global data
- **C++: Arrays of objects and `std::vector<>` are by default initialized sequentially**
  - STL allocators provide an elegant solution

Coding for Data Locality:
Placement of static arrays or arrays of objects

optional

- **Don't forget that constructors tend to touch the data members of an object. Example:**

```
class D {
  double d;
public:
  D(double _d=0.0) throw() : d(_d) {}
  inline D operator+(const D& o) throw() {
    return D(d+o.d);
  }
  inline D operator*(const D& o) throw() {
    return D(d*o.d);
  }
...
};
```

→ **placement problem with**

```
D* array = new D[1000000];
```

Coding for Data Locality:
Parallel first touch for arrays of objects

optional

- **Solution: Provide overloaded `D::operator new[]`**

```
void* D::operator new[](size_t n) {
  char *p = new char[n];    // allocate

  size_t i,j;
#pragma omp parallel for private(j) schedule(...)
  for(i=0; i<n; i += sizeof(D))
    for(j=0; j<sizeof(D); ++j)
      p[i+j] = 0;
  return p;
}

void D::operator delete[](void* p) throw() {
  delete [] static_cast<char*>p;
}
```

parallel first touch

- **Placement of objects is then done automatically by the C++ runtime via “placement new”**

Coding for Data Locality:

NUMA allocator for parallel first touch in `std::vector<>`

optional

```
template <class T> class NUMA_Allocator {
public:
  T* allocate(size_type numObjects, const void
              *localityHint=0) {
    size_type ofs,len = numObjects * sizeof(T);
    void *m = malloc(len);
    char *p = static_cast<char*>(m);
    int i,pages = len >> PAGE_BITS;
#pragma omp parallel for schedule(static) private(ofs)
    for(i=0; i<pages; ++i) {
      ofs = static_cast<size_t>(i) << PAGE_BITS;
      p[ofs]=0;
    }
    return static_cast<pointer>(m);
  }
...
};
```

Application:

```
vector<double,NUMA_Allocator<double> > x(10000000)
```

Diagnosing bad locality

- **If your code is cache bound, you might not notice any locality problems**
- **Otherwise, bad locality limits scalability (whenever a ccNUMA node boundary is crossed)**
  - Just an indication, not a proof yet
- **Running with `numactl --interleave` might give you a hint**
  - See later
- **Consider using performance counters**
  - LIKWID-perfctr can be used to measure nonlocal memory accesses
  - Example for Intel Westmere dual-socket system (IvyBridge, 2x10-core):

```
likwid-perfctr -g NUMA –C M0:0-4@M1:0-4 ./a.out
```

Using performance counters for diagnosing bad ccNUMA access locality

- Intel Ivy Bridge EP node (running 2x5 threads): measure NUMA traffic per core

```
likwid-perfctr -g NUMA –C M0:0-4@M1:0-4 ./a.out
```

- Summary output:

```
+--------------------------------------+-------------+-------------+-------------+--------------+
|                Metric                |     Sum     |     Min     |     Max     |     Avg      |
+--------------------------------------+-------------+-------------+-------------+--------------+
|       Runtime (RDTSC) [s] STAT       |  4.050483   |  0.4050483  |  0.4050483  |   0.4050483  |
|       Runtime unhalted [s] STAT      |   3.03537   |  0.3026072  |  0.3043367  |   0.303537   |
|           Clock [MHz] STAT           |  32996.94   |   3299.692  |   3299.696  |   3299.694   |
|               CPI STAT               |   40.3212   |   3.702072  |   4.244213  |    4.03212   |
|  Local DRAM data volume [GByte] STAT | 7.752933632 | 0.735579264 | 0.823551488 | 0.7752933632 |
|  Local DRAM bandwidth [MByte/s] STAT |  19140.761  |   1816.028  |   2033.218  |   1914.0761  |
| Remote DRAM data volume [GByte] STAT | 9.16628352  |  0.86682464 | 0.957811776 |  0.916628352 |
| Remote DRAM bandwidth [MByte/s] STAT |  22630.098  |   2140.052  |   2364.685  |   2263.0098  |
|    Memory data volume [GByte] STAT   | 16.919217152| 1.690376128 |  1.69339104 | 1.6919217152 |
|    Memory bandwidth [MByte/s] STAT   |  41770.861  |   4173.27   |   4180.714  |   4177.0861  |
+--------------------------------------+-------------+-------------+-------------+--------------+
```

About half of the overall memory traffic is caused by remote domain!

- Caveat: NUMA metrics vary strongly between CPU models

If all fails…

optional

- **Even if all placement rules have been carefully observed, you may still see nonlocal memory traffic. Reasons?**
  - Program has erratic access patters → may still achieve some access parallelism (see later)
  - OS has filled memory with buffer cache data:

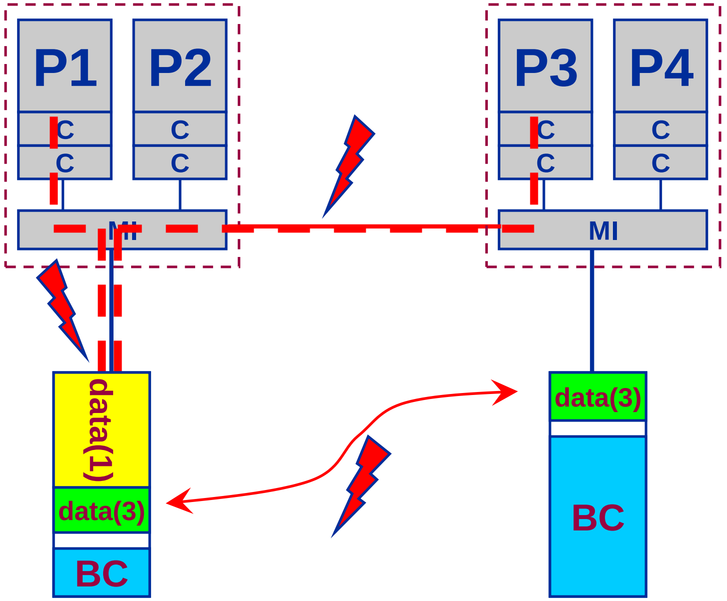
```
# numactl --hardware    # idle node!
available: 2 nodes (0-1)
node 0 size: 2047 MB
node 0 free: 906 MB
node 1 size: 1935 MB
node 1 free: 1798 MB
```

```
top - 14:18:25 up 92 days,  6:07,  2 users,  load average: 0.00, 0.02, 0.00
Mem:   4065564k total,  1149400k used,  2716164k free,    43388k buffers
Swap:  2104504k total,     2656k used,  2101848k free,  1038412k cached
```

ccNUMA problems beyond first touch:
Buffer cache

optional

- **OS uses part of main memory for disk buffer (FS) cache**
  - If FS cache fills part of memory, apps will probably allocate from foreign domains
  - → non-local access!
  - "sync" is not sufficient to drop buffer cache blocks



- **Remedies**
  - Drop FS cache pages after user job has run (admin's job)
    - seems to be automatic after aprun has finished on Crays
  - User can run "sweeper" code that allocates and touches all physical memory before starting the real application
  - `numactl` tool or `aprun` can force local allocation (where applicable)
  - Linux: There is no way to limit the buffer cache size in standard kernels

ccNUMA problems beyond first touch:
Buffer cache

optional

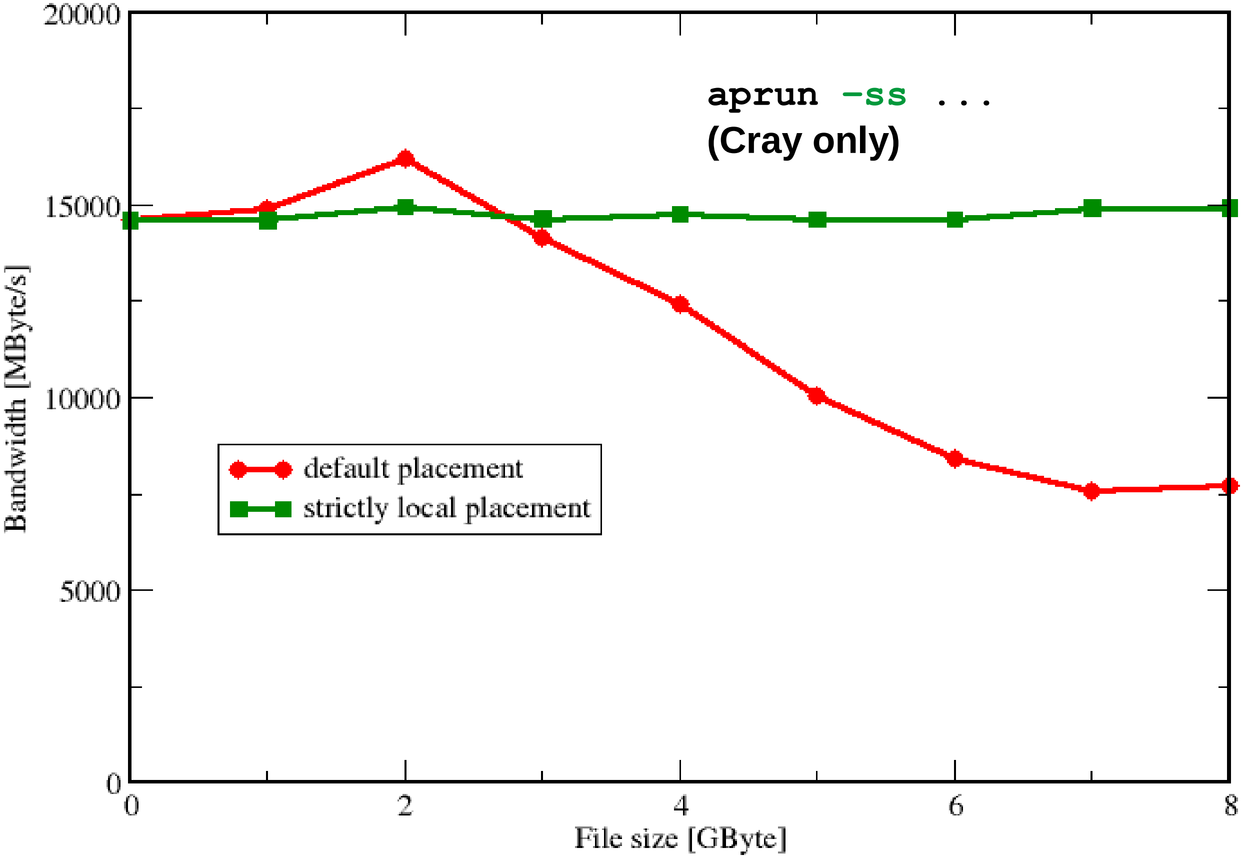
Real-world example: ccNUMA and the Linux buffer cache

Benchmark:

1. Write a file of some size from LD0 to disk
2. Perform bandwidth benchmark using all cores in LD0 and maximum memory installed in LD0

Result: By default, Buffer cache is given priority over local page placement

→ restrict to local domain if possible!

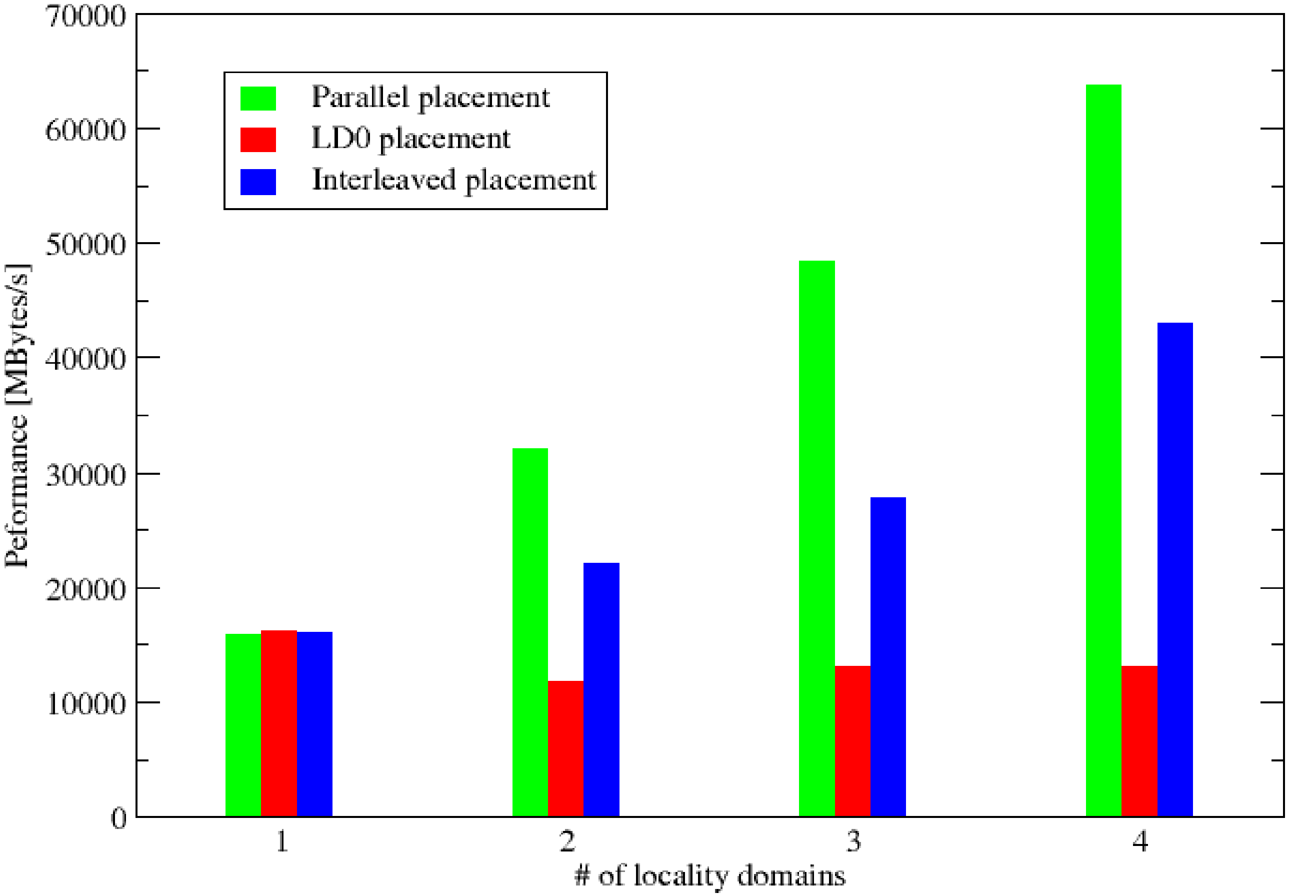


```
aprun -ss ...
```

(Cray only)

The curse and blessing of interleaved placement:
OpenMP STREAM on a Cray XE6 Interlagos node

- **Parallel init: Correct parallel initialization**
- **LD0: Force data into LD0 via `numactl –m 0`**
- **Interleaved: `numactl --interleave <LD range>`**



Summary on ccNUMA issues

- **Identify the problem**
  - Is ccNUMA an issue in your code?
  - Simple test: run with `numactl --interleave`

- **Apply first-touch placement**
  - Look at initialization loops
  - Consider loop lengths and static scheduling
  - C++ and global/static objects may require special care

- **If dynamic scheduling cannot be avoided**
  - Consider round-robin placement / page interleaving

- **Buffer cache may impact proper placement**
  - Kick your admins
  - or apply sweeper code
  - If available, use runtime options to force local placement