# Application behaviour and Numaplace

Daniel J Blueman

Principal Software Engineer, Numascale

March 10, 2016

# Introduction

- Exclusive goals
    - Want desktop-like flexibility and freedom on larger systems
    - Want all the performance
- Large-SMP systems have much higher scheduling overhead
    - scheduling decision complexity is exponential with core count
    - takes away some* performance
- * Depends on how much sleeping on semaphores vs compute-bound the workload is

# Existing approach

- Manipulate environment
  - Look! Cores 64 to 191 are free

    ```
    $ export OMP_NUM_THREADS=128
    $ export OMP_PLACES={64-191}
    $ export OMP_PROC_BIND=true
    $ ./benchmark
    ```

- Caveats
  - Wait, it's running slow now
    - Someone else is running on those cores
  - How can I tell which cores are available?
    - No robust mechanism. Don't even think about htop...
- OMP_PLACES what?
  - Sorry, needs OpenMP 4.0

# Making life easier

- Abstracts guesswork of cores
  - so you don't have to
- Gives desktop-like scheduling latency
- Transparent to OpenMP or pthreads application
- Automatically isolates applications from each other
- Detects NUMA topology and optimises core placement
- Published at:
  - `https://resources.numascale.com/numaplace/`
- Source at:
  - `https://github.com/numascale/nc-utils/tree/master/os/numaplace`
- WIP to integrate into numactl package, so is OOB

# Options

```
usage: numaplace [-atvVdp] [-c <cores>] cmd [args ...]
        -a, --no-allocator      don't use NUMA aware memory
        -c, --cores             set number of cores adverti
        -d, --debug             show internal information
        -p, --parent            don't pin parent task
        -t, --no-thp            disable Transparent Huge Pa
        -v, --verbose           show cores allocated
        -V, --version           show version
```

# Example

```
$ numaplace --cores 64 ./cg.C.x
 NAS Parallel Benchmarks (NPB3.3-OMP) - CG Benchmark
...
 Number of available threads:    64
...
       75       0.90260466198027E-15    28.9736055928455
 Benchmark completed
 VERIFICATION SUCCESSFUL
...
 Mop/s total     =                   903.41
```

# Improvements

- Wait, but only 903 MFLOPS?
- Maybe the application doesn't interact well with transparent hugepages...

```
$ numaplace --cores 64 --no-thp ./cg.C.x
 NAS Parallel Benchmarks (NPB3.3-OMP) - CG Benchmark
...
 Number of available threads:    64
...
       75       0.90260466198027E-15    28.9736055928455
 Benchmark completed
 VERIFICATION SUCCESSFUL
...
 Mop/s total     =                 3230.63
```

- 3.2 GFLOPS...much better!

# Roadmap

- Stride allocation
  - Prevents FPU sharing
  - May give more per-thread memory bandwidth
- Automatic core count
  - Will check how many cores aren't used
  - Configurable default limit
    - so multiple users can share a system effectively
- Transparent-hugepage blacklist
  - Disables THP for applications which are known to behave poorly

# Thankyou

- Do drop me a note at daniel@numascale.com
- Feedback, issues or requests welcome