Profiling & Parallelization

Lecture 21

Dr. Colin Rundel

Profiling & Benchmarking

profvis demo

```r
n = 1e6
d = tibble(
  x1 = rt(n, df = 3),
  x2 = rt(n, df = 3),
  x3 = rt(n, df = 3),
  x4 = rt(n, df = 3),
  x5 = rt(n, df = 3),
) |>
  mutate(y = -2*x1 - 1*x2 + 0*x3 + 1*x4 + 2*x5 + rnorm(n))
```

```r
profvis::profvis(lm(y~., data=d))
```

Benchmarking - bench

```
d = tibble(
  x = runif(10000),
  y = runif(10000)
)

(b = bench::mark(
  d[d$x > 0.5, ],
  d[which(d$x > 0.5), ],
  subset(d, x > 0.5),
  filter(d, x > 0.5)
))
```

```
# A tibble: 4 × 6
  expression                    min   median `itr/sec` mem_alloc `gc/sec`
  <bch:expr>               <bch:tm> <bch:tm>     <dbl> <bch:byt>    <dbl>
1 d[d$x > 0.5, ]              128µs    140µs     7002.  239.84KB     19.4
2 d[which(d$x > 0.5), ]       137µs    152µs     6388.  271.49KB     33.5
3 subset(d, x > 0.5)          167µs    195µs     4940.  288.85KB     26.3
4 filter(d, x > 0.5)          379µs    432µs     2167.     1.48MB     35.4
```

Larger n

```r
d = tibble(
  x = runif(1e6),
  y = runif(1e6)
)

(b = bench::mark(
  d[d$x > 0.5, ],
  d[which(d$x > 0.5), ],
  subset(d, x > 0.5),
  filter(d, x > 0.5)
))
```

```
# A tibble: 4 × 6
  expression                   min   median `itr/sec` mem_alloc `gc/sec`
  <bch:expr>              <bch:tm> <bch:tm>     <dbl> <bch:byt>    <dbl>
1 d[d$x > 0.5, ]              12ms   12.4ms      80.6    13.4MB     60.4
2 d[which(d$x > 0.5), ]     13.3ms   13.6ms      73.6    24.8MB    210. 
3 subset(d, x > 0.5)        17.7ms   17.9ms      55.7    24.8MB     65.0
4 filter(d, x > 0.5)        13.5ms   13.9ms      71.9    24.8MB     56.8
```

bench - relative results

```r
summary(b, relative=TRUE)
```

```
# A tibble: 4 × 6
  expression                    min median `itr/sec` mem_alloc `gc/sec`
  <bch:expr>                  <dbl>  <dbl>     <dbl>     <dbl>    <dbl>
1 d[d$x > 0.5, ]               1      1         1.45      1        1.06
2 d[which(d$x > 0.5), ]  1.11   1.09      1.32      1.86     3.70
3 subset(d, x > 0.5)         1.47   1.45      1         1.86     1.14
4 filter(d, x > 0.5)         1.12   1.12      1.29      1.86     1
```

`t.test`

> Imagine we have run 1000 experiments (rows), each of which collects data on 50 individuals (columns). The first 25 individuals in each experiment are assigned to group 1 and the rest to group 2.

The goal is to calculate the t-statistic for each experiment comparing group 1 to group 2.

```r
m = 1000
n = 50
X = matrix(
  rnorm(m * n, mean = 10, sd = 3),
  ncol = m
) |>
  as.data.frame() |>
  set_names(paste0("exp", seq_len(m))) |>
  mutate(
    ind = seq_len(n),
    group = rep(1:2, each = n/2)
  ) |>
  as_tibble() |>
  relocate(ind, group)
X
```

```
# A tibble: 50 × 1,002
     ind group  exp1  exp2  exp3  exp4  exp5  exp6
   <int> <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
 1     1     1 12.1  15.0  12.8   5.54 12.0  11.5
 2     2     1  5.61  6.99 14.3   8.47  6.49  5.17
 3     3     1 12.5  11.6  15.7  10.3  14.0   7.52
 4     4     1  8.33  8.39 10.5  15.6   9.94  8.21
 5     5     1 11.1   5.80  8.22  9.48  5.08  2.59
 6     6     1  9.48  6.19  6.05  9.16  8.69 17.0
 7     7     1 11.2  11.8   5.98  6.05  9.70 12.9
 8     8     1  9.56  8.14 14.6  11.4   7.39  9.63
 9     9     1 10.1   8.72 10.3   7.71 11.5  12.1
10    10     1 17.1  11.6   8.38 12.3   7.88  8.59
# ℹ 40 more rows
# ℹ 994 more variables: exp7 <dbl>, exp8 <dbl>,
#   exp9 <dbl>, exp10 <dbl>, exp11 <dbl>,
#   exp12 <dbl>, exp13 <dbl>, exp14 <dbl>,
```

Implementations

```r
ttest_formula = function(X, m) {
  for(i in 1:m) t.test(X[[2+i]] ~ X$group)$stat
}

system.time(ttest_formula(X,m))
```

```
 user  system elapsed 
0.183   0.001   0.186 
```

```r
ttest_for = function(X, m) {
  for(i in 1:m) t.test(X[[2+i]][X$group == 1], X[[2+i]][X$group == 2])$stat
}

system.time(ttest_for(X,m))
```

```
 user  system elapsed 
0.064   0.001   0.066 
```

```r
ttest_apply = function(X) {
  f = function(x, g) {
    t.test(x[g==1], x[g==2])$stat
  }
  apply(X[,-(1:2)], 2, f, X$group)
}

system.time(ttest_apply(X))
```

```
 user  system elapsed
0.053   0.000   0.054
```

Implementations (cont.)

```r
ttest_hand_calc = function(X) {
  f = function(x, grp) {
    t_stat = function(x) {
      m = mean(x)
      n = length(x)
      var = sum((x - m) ^ 2) / (n - 1)

      list(m = m, n = n, var = var)
    }

    g1 = t_stat(x[grp == 1])
    g2 = t_stat(x[grp == 2])

    se_total = sqrt(g1$var / g1$n + g2$var / g2$n)
    (g1$m - g2$m) / se_total
  }

  apply(X[,-(1:2)], 2, f, X$group)
}
```

```
   user  system elapsed 
  0.014   0.000   0.015 
```

Comparison

```r
bench::mark(
  ttest_formula(X, m),
  ttest_for(X, m),
  ttest_apply(X),
  ttest_hand_calc(X),
  check=FALSE
)
```

```
# A tibble: 4 × 6
  expression               min   median `itr/sec` mem_alloc `gc/sec`
  <bch:expr>          <bch:tm> <bch:tm>     <dbl> <bch:byt>    <dbl>
1 ttest_formula(X, m) 197.05ms 199.89ms      5.01    8.24MB     26.7
2 ttest_for(X, m)      68.75ms   69.1ms     14.3     1.91MB     26.8
3 ttest_apply(X)       58.24ms  62.82ms     16.1     3.49MB     26.8
4 ttest_hand_calc(X)    8.61ms   9.21ms     89.6     3.45MB     25.9
```

Parallelization

`parallel`

Part of the base packages in R

- tools for the forking of R processes (some functions do not work on Windows)
- Core functions:
  - `detectCores`
  - `pvec`
  - `mclapply`
  - `mcparallel` & `mccollect`

detectCores

Surprisingly, detects the number of cores of the current system.

```r
detectCores()
```

```
[1] 10
```

pvec

Parallelization of a vectorized function call

```r
system.time(pvec(1:1e7, sqrt, mc.cores = 1))
```

```
 user  system elapsed 
0.016   0.012   0.028 
```

```r
system.time(pvec(1:1e7, sqrt, mc.cores = 4))
```

```
 user  system elapsed 
0.168   0.151   0.249 
```

```r
system.time(pvec(1:1e7, sqrt, mc.cores = 8))
```

```
 user  system elapsed 
0.092   0.164   0.152 
```

```r
system.time(sqrt(1:1e7))
```

```
 user  system elapsed 
0.055   0.018   0.072 
```

pvec - `bench::system_time`

```
bench::system_time(pvec(1:1e7, sqrt, mc.cores = 1))
```

```
process    real
   25ms  24.8ms
```

```
bench::system_time(pvec(1:1e7, sqrt, mc.cores = 4))
```

```
process    real
  150ms   179ms
```

```
bench::system_time(pvec(1:1e7, sqrt, mc.cores = 8))
```

```
process    real
  191ms   223ms
```

```r
bench::system_time(Sys.sleep(.5))
```

```
process    real 
   64µs   497ms 
```

```r
system.time(Sys.sleep(.5))
```

```
   user  system elapsed 
  0.000   0.000   0.505 
```

Cores by size

```
cores = c(1,4,6,8,10)
order = 6:8
f = function(x,y) {
  system.time(
    pvec(1:(10^y), sqrt, mc.cores = x)
  )[3]
}

res = map(
  cores,
  function(x) {
     map_dbl(order, f, x = x)
  }
) |>
  do.call(rbind, args = _)

rownames(res) = paste0(cores," cores")
colnames(res) = paste0("10^",order)

res
```

```
         10^6  10^7  10^8
1 cores  0.004 0.032 0.371
4 cores  0.032 0.145 2.006
6 cores  0.024 0.138 1.367
8 cores  0.033 0.127 1.265
10 cores 0.033 0.147 1.548
```

mclapply

Parallelized version of `lapply`

```
system.time(rnorm(1e7))
```

```
 user  system elapsed 
0.262   0.004   0.266 
```

```
system.time(unlist(mclapply(1:10, function(x) rnorm(1e6), mc.cores = 2)))
```

```
 user  system elapsed 
0.309   0.095   0.268 
```

```
system.time(unlist(mclapply(1:10, function(x) rnorm(1e6), mc.cores = 4)))
```

```
 user  system elapsed 
0.322   0.089   0.161 
```

```
system.time(unlist(mclapply(1:10, function(x) rnorm(1e6), mc.cores = 8)))
```

```
 user  system elapsed 
0.336   0.145   0.172 
```

```
system.time(unlist(mclapply(1:10, function(x) rnorm(1e6), mc.cores = 10)))
```

```
 user  system elapsed 
0.360   0.150   0.179 
```

mcparallel

Asynchronously evaluation of an R expression in a separate process

```r
m = mcparallel(rnorm(1e6))
n = mcparallel(rbeta(1e6,1,1))
o = mcparallel(rgamma(1e6,1,1))
```

```r
str(m)
```

```
List of 2
 $ pid: int 14040
 $ fd : int [1:2] 4 7
 - attr(*, "class")= chr [1:3] "parallelJob" "childProcess" "process"
```

```r
str(n)
```

```
List of 2
 $ pid: int 14041
 $ fd : int [1:2] 5 9
 - attr(*, "class")= chr [1:3] "parallelJob" "childProcess" "process"
```

mccollect

Checks `mcparallel` objects for completion

```r
str(mccollect(list(m,n,o)))
```

```
List of 3
 $ 14040: num [1:1000000] -0.7172 -1.8334 -0.0983 0.953 0.1629 ...
 $ 14041: num [1:1000000] 0.235 0.554 0.301 0.977 0.644 ...
 $ 14042: num [1:1000000] 0.448 0.192 0.342 1.386 0.397 ...
```

mccollect - waiting

```r
p = mcparallel(mean(rnorm(1e5)))
```

```r
mccollect(p, wait = FALSE, 10)
```

```
$`14043`
[1] -0.001305267
```

```r
mccollect(p, wait = FALSE)
```

```
NULL
```

```r
mccollect(p, wait = FALSE)
```

```
NULL
```

doMC & foreach

doMC & foreach

Packages by Revolution Analytics that provides the `foreach` function which is a parallelizable `for` loop (and then some).

- Core functions:
  - `registerDoMC`
  - `foreach`, `%dopar%`, `%do%`

registerDoMC

Primarily used to set the number of cores used by `foreach`, by default uses `options("cores")` or half the number of cores found by `detectCores` from the parallel package.

```r
options("cores")
```

```
$cores
NULL
```

```r
detectCores()
```

```
[1] 10
```

```r
getDoParWorkers()
```

```
[1] 1
```

```r
registerDoMC(4)
getDoParWorkers()
```

```
[1] 4
```

`foreach`

A slightly more powerful version of base `for` loops (think `for` with an `lapply` flavor). Combined with `%do%` or `%dopar%` for single or multicore execution.

```r
for(i in 1:10) {
  sqrt(i)
}

foreach(i = 1:5) %do% {
  sqrt(i)
}
```

```
[[1]]
[1] 1

[[2]]
[1] 1.414214

[[3]]
[1] 1.732051
```

```
[[4]]
[1] 2
```

`foreach` - iterators

`foreach` can iterate across more than one value, but it doesn't do length coercion

```
foreach(i = 1:5, j = 1:5) %
  sqrt(i^2+j^2)
}
```

```
[[1]]
[1] 1.414214

[[2]]
[1] 2.828427

[[3]]
[1] 4.242641

[[4]]
[1] 5.656854

[[5]]
```

```
foreach(i = 1:5, j = 1:2) %
  sqrt(i^2+j^2)
}
```

```
[[1]]
[1] 1.414214

[[2]]
[1] 2.828427
```

foreach - combining results

```r
foreach(i = 1:5, .combine='c') %do% {
  sqrt(i)
}
```

```
[1] 1.000000 1.414214 1.732051 2.000000 2.236068
```

```r
foreach(i = 1:5, .combine='cbind') %do% {
  sqrt(i)
}
```

```
     result.1 result.2 result.3 result.4 result.5
[1,]        1 1.414214 1.732051        2 2.236068
```

```r
foreach(i = 1:5, .combine='+') %do% {
  sqrt(i)
}
```

```
[1] 8.382332
```

`foreach` - parallelization

Swapping out `%do%` for `%dopar%` will use the parallel backend.

```
registerDoMC(4)
system.time(foreach(i = 1:10) %dopar% mean(rnorm(1e6)))
```

```
   user  system elapsed 
  0.298   0.028   0.109 
```

```
registerDoMC(8)
system.time(foreach(i = 1:10) %dopar% mean(rnorm(1e6)))
```

```
   user  system elapsed 
  0.302   0.032   0.076 
```

```
registerDoMC(10)
system.time(foreach(i = 1:10) %dopar% mean(rnorm(1e6)))
```

```
   user  system elapsed 
  0.325   0.042   0.069 
```

furrr / future

```r
system.time( purrr::map(c(1,1,1), Sys.sleep) )
```

```
   user  system elapsed 
  0.000   0.000   3.012 
```

```r
system.time( furrr::future_map(c(1,1,1), Sys.sleep) )
```

```
   user  system elapsed 
  0.053   0.008   3.097 
```

```r
future::plan(future::multisession) # See also future::multicore
system.time( furrr::future_map(c(1,1,1), Sys.sleep) )
```

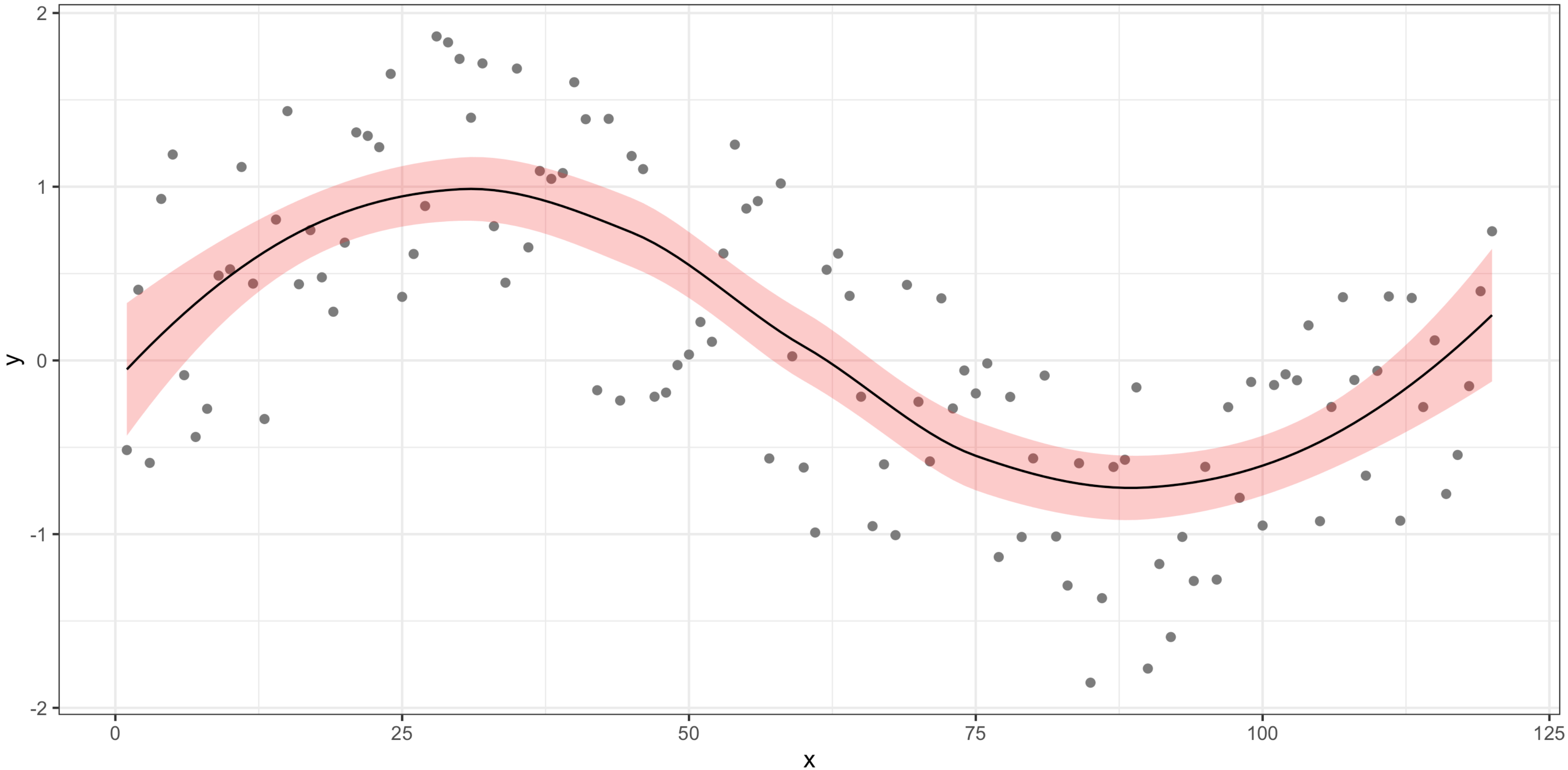
```
   user  system elapsed 
  0.206   0.007   1.451 
```

Example - Bootstraping

Bootstrapping is a resampling scheme where the original data is repeatedly reconstructed by taking a samples of size *n* (with replacement) from the original data, and using that to repeat an analysis procedure of interest. Below is an example of fitting a local regression (`loess`) to some synthetic data, we will construct a bootstrap prediction interval for this model.

```
set.seed(3212016)
d = data.frame(x = 1:120) |>
    mutate(y = sin(2*pi*x/120) + runif(length(x),-1,1))

l = loess(y ~ x, data=d)
p = predict(l, se=TRUE)

d = d |> mutate(
  pred_y = p$fit,
  pred_y_se = p$se.fit
)
```

```r
ggplot(d, aes(x,y)) +
  geom_point(color="gray50") +
  geom_ribbon(
    aes(ymin = pred_y - 1.96 * pred_y_se,
        ymax = pred_y + 1.96 * pred_y_se),
    fill="red", alpha=0.25
  ) +
  geom_line(aes(y=pred_y)) +
  theme_bw()
```



2
1
0
-1
-2
y
0
25
50
75
100
125
x

Bootstraping Demo

What to use when?

Optimal use of parallelization / multiple cores is hard, there isn't one best solution

- Don't underestimate the overhead cost
- Experimentation is key
- Measure it or it didn't happen
- Be aware of the trade off between developer time and run time

BLAS and LAPACK

Statistics and Linear Algebra

An awful lot of statistics is at its core linear algebra.

For example:

- Linear regession models, find

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

- Principle component analysis
  - Find $T = XW$ where W is a matrix whose columns are the eigenvectors of $X^T X$.
  - Often solved via SVD - Let $X = U \Sigma W^T$ then $T = U \Sigma$.

Numerical Linear Algebra

Not unique to Statistics, these are the type of problems that come up across all areas of numerical computing.

- Numerical linear algebra $\neq$ mathematical linear algebra
- Efficiency and stability of numerical algorithms matter
  - Designing and implementing these algorithms is hard
- Don't reinvent the wheel - common core linear algebra tools (well defined API)

BLAS and LAPACK

Low level algorithms for common linear algebra operations

BLAS

- **B**asic **L**inear **A**lgebra **S**ubprograms
- Copying, scaling, multiplying vectors and matrices
- Origins go back to 1979, written in Fortran

LAPACK

- **L**inear **A**lgebra **Pack**age
- Higher level functionality building on BLAS.
- Linear solvers, eigenvalues, and matrix decompositions
- Origins go back to 1992, mostly Fortran (expanded on LINPACK, EISPACK)

Modern variants?

Most default BLAS and LAPACK implementations (like R's defaults) are somewhat dated

- Written in Fortran and designed for a single cpu core
- Certain (potentially non-optimal) hard coded defaults (e.g. block size).

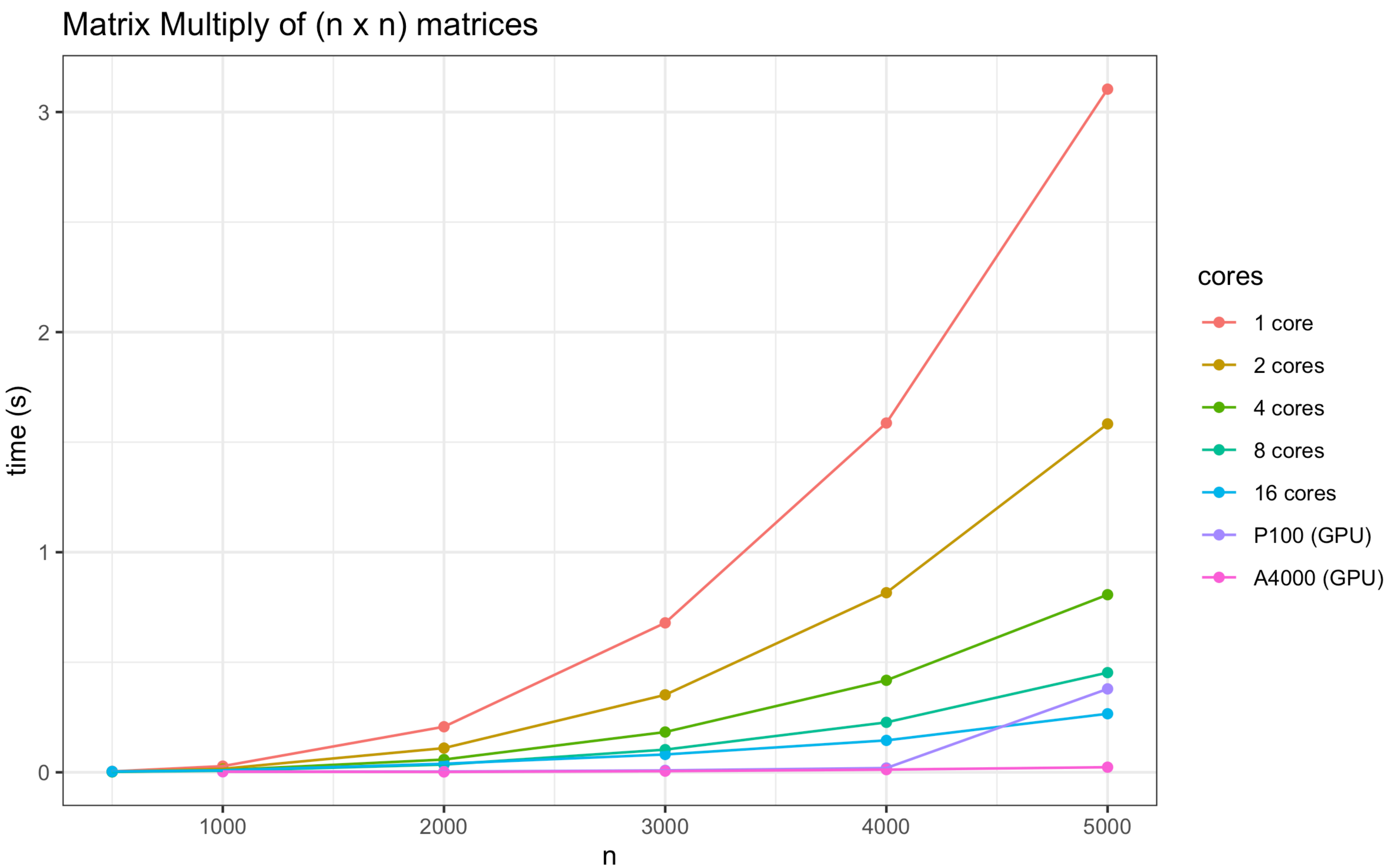
Multithreaded alternatives:

- ATLAS - Automatically Tuned Linear Algebra Software
- OpenBLAS - fork of GotoBLAS from TACC at UTexas
- Intel MKL - Math Kernel Library, part of Intel's commercial compiler tools
- cuBLAS / Magma - GPU libraries from Nvidia and UTK respectively
- Accelerate / vecLib - Apple's framework for GPU and multicore computing

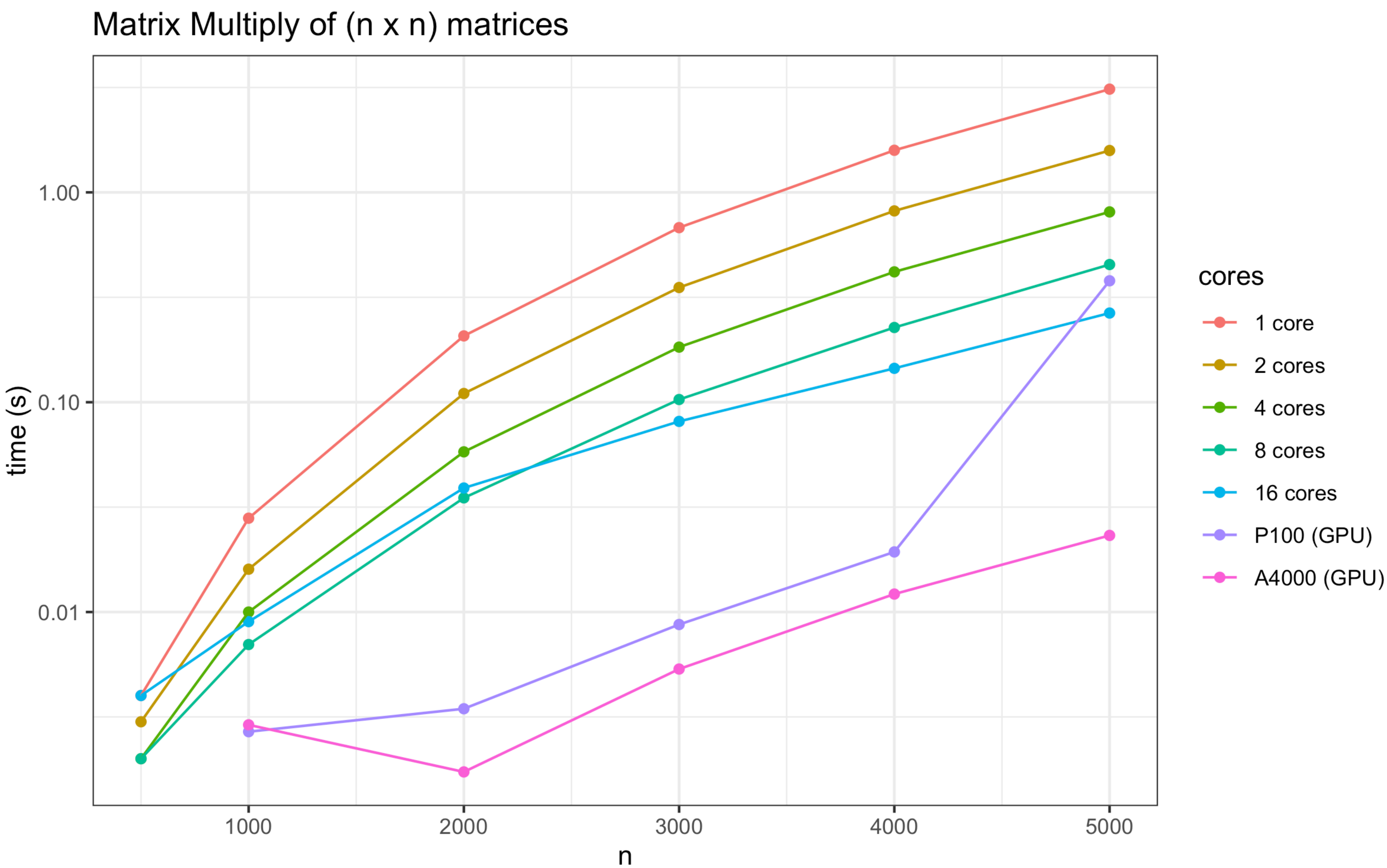
OpenBLAS Matrix Multiply Performance

```r
x=matrix(runif(5000^2),ncol=5000)

sizes = c(100,500,1000,2000,3000,4000,5000)
cores = c(1,2,4,8,16)

sapply(
  cores,
  function(n_cores)
  {
    flexiblas::flexiblas_set_num_threads(n_cores)
    sapply(
      sizes,
      function(s)
      {
        y = x[1:s,1:s]
        system.time(y %*% y)[3]
      }
    )
  }
)
```

n	1 core	2 cores	4 cores	8 cores	16 cores
100	0.000	0.000	0.000	0.000	0.000
500	0.004	0.003	0.002	0.002	0.004
1000	0.028	0.016	0.010	0.007	0.009
2000	0.207	0.110	0.058	0.035	0.039
3000	0.679	0.352	0.183	0.103	0.081
4000	1.587	0.816	0.418	0.227	0.145
5000	3.104	1.583	0.807	0.453	0.266



Matrix Multiply of (n x n) matrices
time (s)
3
2
1
0
1000
2000
3000
4000
5000
n
cores
1 core
2 cores
4 cores
8 cores
16 cores
P100 (GPU)
A4000 (GPU)


Matrix Multiply of (n x n) matrices
time (s)
1.00
0.10
0.01
1000
2000
3000
4000
5000
n
cores
1 core
2 cores
4 cores
8 cores
16 cores
P100 (GPU)
A4000 (GPU)