Rcpp - Part 1

Jon Meek
meekjt (at) gmail.com
meekj (at) ieee.org

https://github.com/meekj/TRU-July2016

23-July-2016 / Trenton R Users

Warning - Prepared on Short Notice!

- I agreed to do this last night
- I did not decide to make slides until four hours ago
- So, coherence and completeness will be limited

What is Rcpp? Why use it?

- Easy way to incorporate C++ into R code
- Use Julia for speed? Dirk Eddelbuettel says use Rcpp
- Loops in R are slow
  - vectorize if possible
  - if not possible use Rcpp
- Integrate C/C++ libraries into R
- Using R + C++ is similar to how I paired FORTRAN + Assembly and Pascal + Assembly in the far past

My First Rcpp Application

- Compute estimated bandwidth or concurrent sessions from network device log files
  - Millions, or a billion, records
  - Use session start / stop times
  - Distribute total bytes,active sessions, or unique users, across one second bins

Sample Code - Concurrent Session Counts

```
library(Rcpp)

cppFunction('NumericVector ccsCalc(int outlen, NumericVector Second, NumericVector iDuration) {
   NumericVector ccs(outlen);
   int k, j;
   int n = Second.size();

   for(int i = 0; i < n; ++i) {
     if (iDuration[i] == 0) {
       ccs[Second[i]]++;
     } else {
       k = Second[i] + iDuration[i];
       if (k >= outlen) {k = outlen - 1;}
         for (j = Second[i]; j <= k; j++) {
         ccs[j]++;
       }
     }
   }
   return ccs;
}')

events <- read_delim(f, delim = ' ', col_types = list(Time = col_datetime('%Y-%m-%dT%H:%M:%S')), progress = int
events$StartTime <- events$Time - events$Duration # Fast   0.071   0.048   0.119

# Number of one second bins
timerange_s <- as.integer(difftime(max(events$StartTime), min(events$StartTime), units = 'sec')) + 1

# Index, 0 is first channel for C++, fast, but probably eliminate later
events$Second <- as.integer(difftime(events$StartTime, MinTime, units = 'sec'))

system.time( ccs2 <- ccsCalc(timerange_s, events$Second, events$Duration) )
```

Compute Concurrent Sessions - Results

- 25 million log events
- Two nested loops
- Naive R: about 1 hour
- Vectorize inner loop: 15.9 minutes
- Rcpp: 1.4 seconds !

libpcapR Package

- Load network packet capture into a data frame using libpcap
- https://github.com/meekj/libpcapR
- Still need automated tests, and some users
- Requires libpcap-dev package to be installed.
- Probably only works on Linux and Mac
- Odd relibility issues. Build sometimes fails on a first try, then is OK
- Does not appear on the list of GitHub R packages

Rcpp Resources

- Start here: Advanced R Programming by Hadley Wickham: `http://adv-r.had.co.nz/`
- Full book: Seamless R and C++ Integration with Rcpp by Dirk Eddelbuettel
- Google $\rightarrow$ Stackoverflow are your friends, as expected

Other R Related Projects

- Rcpp log file parser to load data frames
- Estimated bandwidth usage from log files using Rcpp
- Use Perl Net::Netmask data to identify network that contains an IP address (currently vaporware!) - might use Rcpp
- MACaddrR - Replaces manufacturer portion of hex address with abbreviated mfg name
- iperf network stress testing tools (Perl, C++, analysis in R)
- Weather / tidal data analysis for Chesapeake Bay and Delaware River & Bay
- Vibration analysis / guitar tuner in R
- OscilloscopeR - download and process data, possibly a Shiny app

Other Things

- An update to my May talk is coming
  - littler is not needed for the scripts
  - Rscript + docopt works fine
- https://github.com/meekj/TRU-May2016
- UseR 2016 Conference videos
  - https://channel9.msdn.com/Events/useR-international-R-User-conference/useR2016
  - Thanks Microsoft !