parallel processing in sas

November 4, 2019

1 parallel processing in sas

torsten dahlén

2 outline parallel processing

- background, terminology
- in sas, in general
- method: “poor mans” parallel processing at kep in sas
- then i think jing has some quiz to increase “knowledge retention”

2.1 recap ‘parallel processing’ from helga

- to perform independent tasks in parallel rather than consecutively

```
dateset A --> t_1 --> t_2 --> ... -> t_n --> Output
```

- in parallel

```
dataset A --> t_1 --> OutputB1
dataset A --> t_2 --> OutputB2
dataset A --> t_n --> OutputB3
                                  ---> Merge OutputB:
```

2.2 basics: threads vs cpus vs parallel processing

- `threads` = number of concurrent processes. single cpu can handle multiple tasks at the same time but -> depends on clock-speed -> increased heat production -> inefficient processing.
- `cpu cores` = threading on different cores -> less energy expenditure -> more efficient processing
- `parallel processing` = break down tasks to smaller units to be processed in parallel.

2.2.1 tasks are either

- `i/o bound` = the read/write speeds are limiting the task, solution = get multiple ssd's that sas can "thread" data to
- `cpu/process bound` = the cpu utilization is limiting the task, solution = threading

2.2.2 gains: amdahl's law

- the decrease in computation time is limited by tasks that cannot be parallelized
- thus, slighlty decreasing returns with increased number of cpus because of tasks that cannot be parallelized

2.2.3 do i need to parallelize my code?

- `options fullstimer;` = get full log information on time and memory consumption to identify possible shortcomings in your resources.
- run task manager to se resource usage.
- ...but simply do i perform time-consuming tasks on my datasets that are independet?

2.3 how many cpus do i have according to sas?

```
%put (&sysncpu);
(4)
```

2.4 parallel processing in sas

- sas base has a few built-in proc's that offer the use of multiple threads.

```
proc means
proc report
proc sort
proc summary
proc tabulate
proc sql
```

- options: `CPUCOUNT=` specifies how many CPUs can be used, can be set to numeric value or `ACTUAL` `THREAD|NOTHREADS` = controls whether to use threads.

If the `THREADS` system option is set to `NOTHREADS`, the `CPUCOUNT=` option has no effect.

2.5 parallel processing in sas ii

- in sas stat the following procedures have abilities to use multiple threads

```
proc adaptivereg
proc fmm
proc glm
proc glmselect
proc loess
proc mixed
proc quantlife
proc quantreg
proc quantselect
proc rubustreg
```

2.6 parallel processing in sas

- example

```
options threads cpucount=4;

proc sql;
create table blaha as select
a.* from big_data_a a inner join big_data_b b
on a.id=b.id
where a.id in (select distinct id from super_large_data where huge_list_of_obs between 3222 an
order by a.id;
quit;
```

2.7 how can i do parallel processing in sas at kep?

- *"poor mans"* version
- local sas/kep-sas-servers (*no sas connect with remote submit*)
- we can use `systask`

2.8 what is `systask` ?

- = launch and handle shell tasks from within sas (r has `shell()` function)
- **very simple: launch concurrent sas sessions from within sas and tell sas when they are completed**

can be used for pretty much anything: use `systask` to launch spotify, use instead of `proc iml` (needs kep_admin password to configure) to run r from sas

2.9 framework for parallel processing

```
libname temp 'path';
```

- move dataset that are used to your temp-path
- generate `.sas` to be processed in parallel and where the same libname is called to access same path

- run a `systask` eg

```
systask command "sas H:\temp\a1.sas" nowait taskname=a1;
systask command "sas H:\temp\a2.sas" nowait taskname=a2;
waitfor _all_ a1 a2;
```

- do whatever with resulting datasets
- log files will be created and saved as `a1.log` and `a2.log` in the same dir as `.sas` files.

2.10 live example using the code challange

- read dataset from `.csv`
- very simplistic case-control study, calculate or for exposure to pneumonia and/or mononucleosis on the risk of ms.
- what parts of this could be parallelized?

import csv? create studybase (casecontrols and diagnosis information) ? categorize into age groups? sum groups? dichotomize exposures? running the logistic regression? calculate age-specific means?