

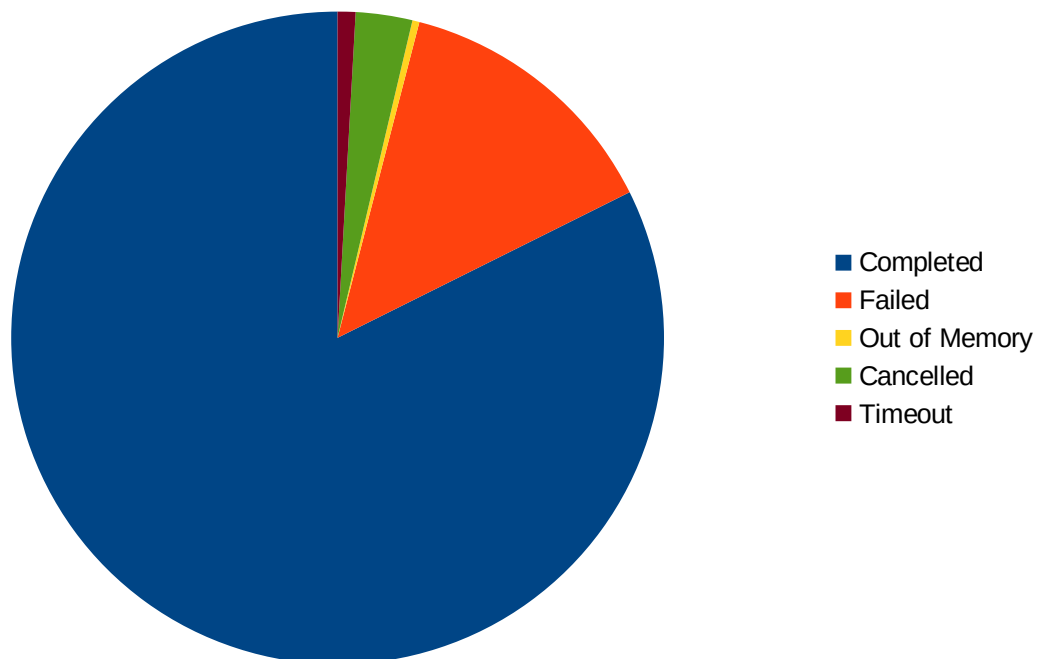
2020 usage in numbers

Due to circumstances 2020 saw a fluctuating number of available CPUs – some of the hardware already retired was brought back online for HPC duties. I managed to get a good estimation of average number of online CPUs throughout the year and used that to calculate averages.

Jobs were asking for ~3376 CPU years of time, which is about 35% more compared to 2019. This is only 35.75% of available CPU time and is less than half of relative allocation last year. Yet we had occasions in first half of the year when jobs were waiting for resources uncomfortably long. If this pattern continues in 2021, it might become necessary to implement some form of scheduling in front of scheduler – organizing larger workloads in projects and giving projects time slots to access HPC resources. Lets wait and see.

In terms of CPU resources used about 1975 CPU years were used, growth of about 23% from last year. This is 20.91% of available CPU resources, which is less than allocation increase, indicating that jobs were more greedy and/or less efficient.

Overall we processed 24.1M jobs, of which 19.5M were completed successfully, 3.2M failed, 665K were canceled, 211K timed out and 83K ran out of memory.



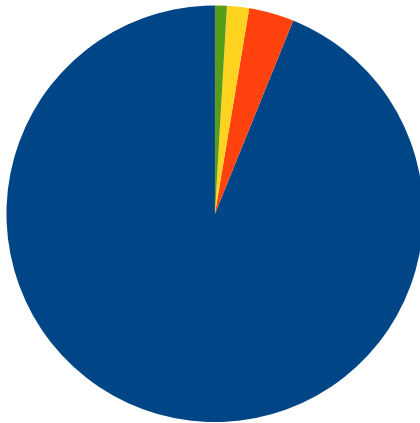
Compared to 2019 we had 1.3M less jobs, but they still consumed more CPU time, which is a good trend – schedulers prefer less but longer running jobs. Lets hope this trend continues.

One unexpected sight here is the large number of failed jobs. I didn't dig into details what those jobs were and what was the cause. For now I naively assume this is an exception.

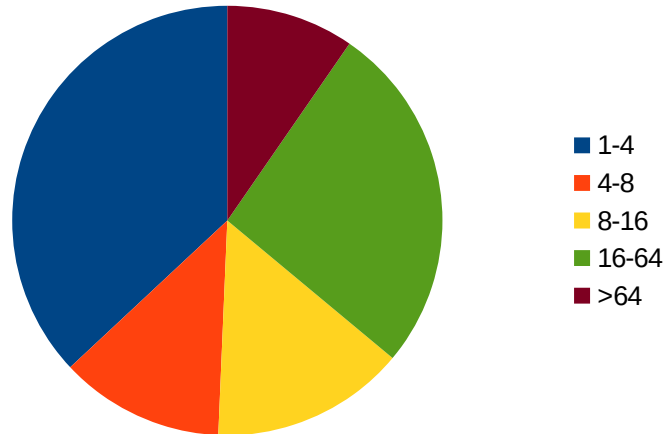
The rest of the analysis is on the completed jobs only. I also looked if adding failed jobs changes things but saw no meaningful changes.

Jobs by size

By number



By CPU time consumed

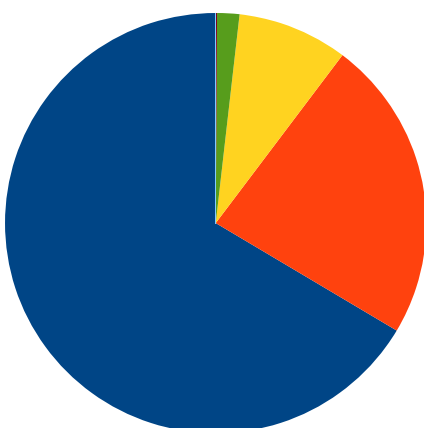


We continue to have large majority of small, “laptop sized” jobs. Their share is even slightly larger than last year. Interestingly, share of “workstation sized” (8-16 cores) jobs decreased the most from last year.

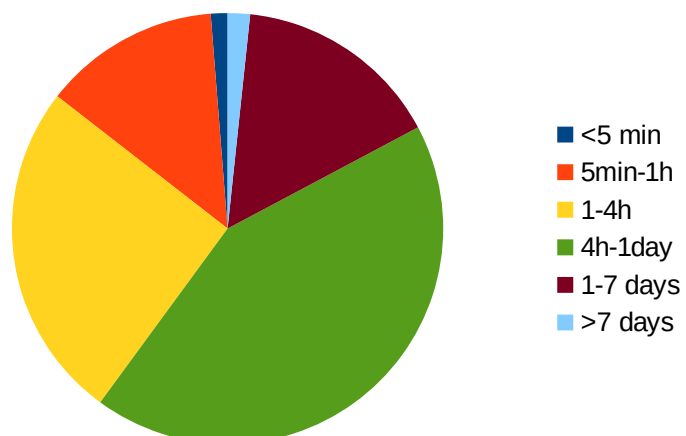
Time consumed shows very slight changes from 2019: small jobs have about the same share, “desktop sized” jobs show a decrease, “workstation sized” jobs also have about the same share, “server sized” have the highest increase of all and “cluster sized” show a decrease.

Jobs by duration

By number



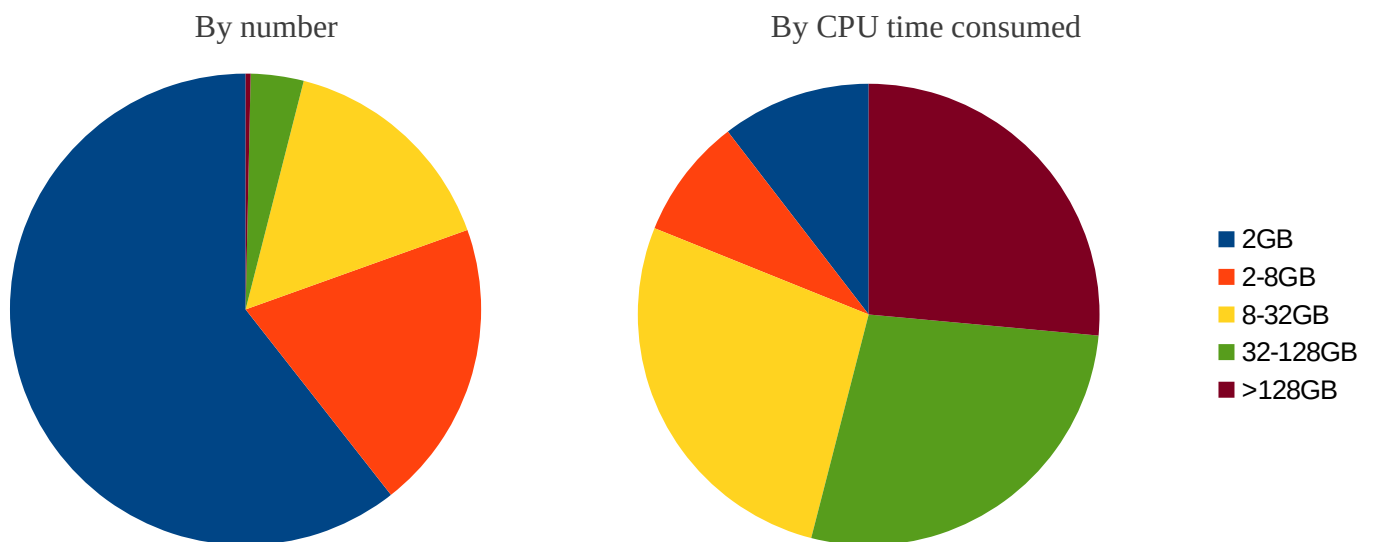
By CPU time consumed



We still have a majority of jobs that are very short. But there’s a noticeable increase in number of jobs running between 1h and 4h compared to last year, which is good.

Time consumed shows similar improvements – decreases for jobs up to 1h, increase for 1-4h, while longer jobs occupied very similar chunks of time.

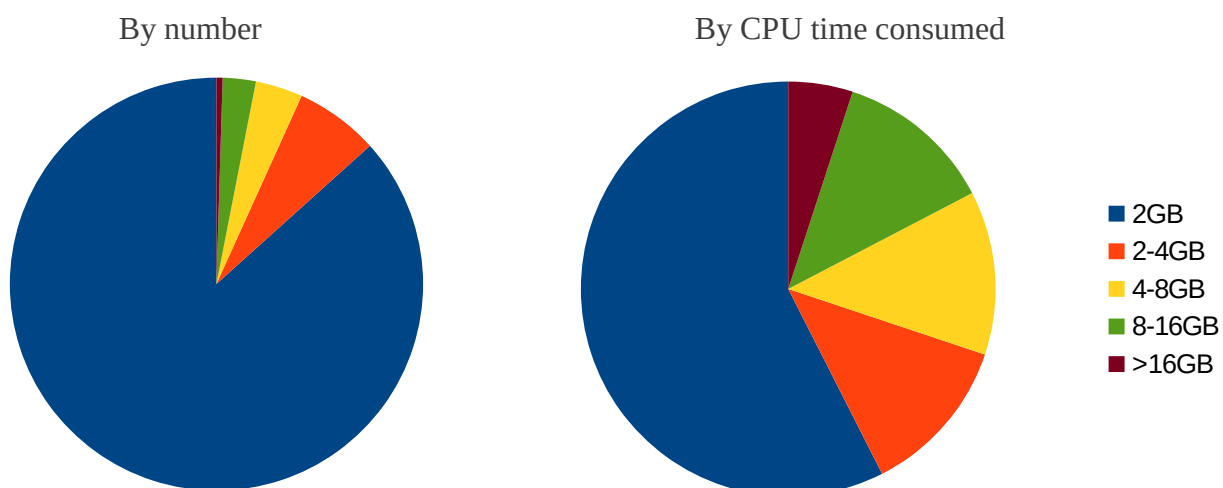
Jobs by memory allocation



Here we have the first major change from last year – there was a large increase in the number of jobs asking for 2GB or less and large matching decrease of jobs asking for 2-8GB of memory. The rest remained about the same. This is a good indication that more attention is being directed towards actual job requirements.

On the time consumed side, there was a decrease in small jobs and increase in largest jobs while middle three remained at about the same share.

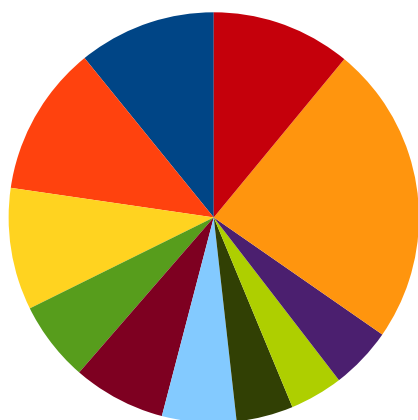
Jobs by memory usage per core



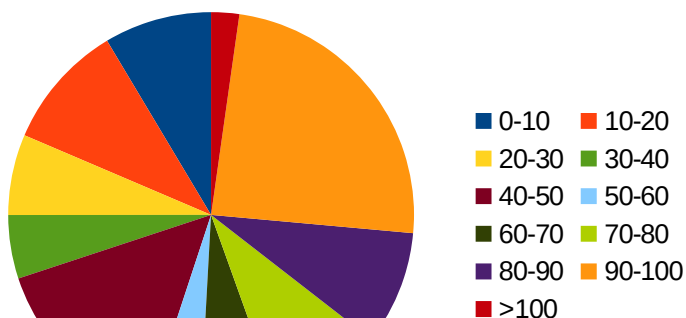
Interesting changes here too – while jobs with up to 2GB per core still dominate, their share has gone down noticeably from last year. All four other kind of jobs show increase. I'd like to know if this is an indication of some larger trend in bioinformatics or just a reflection of the kind of analysis that were ran last year since this greatly affects the hardware configurations we're buying.

CPU efficiency

By number



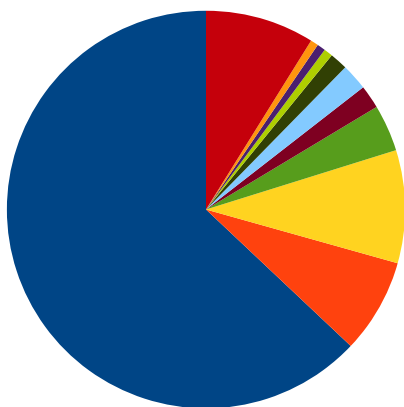
By CPU time consumed



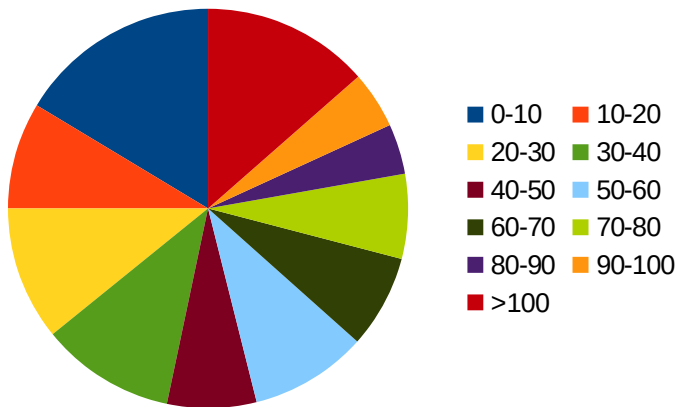
This is where we did much worse than in 2019. About 2/3 of the jobs show less than 90% cpu efficiently while last year this number was only 1/3. By cpu time this year we have only slightly over one fourth of good jobs, while last year this number was about one third.

Jobs by memory efficiency

By number



By CPU time consumed



Situation worsened here too, at both ends: chunk of jobs using less than 10% of allocated memory is about 10% larger compared to last year, but even more worryingly, share of jobs that used over 100% of allocated memory (therefore swapping and running slowly) is about three times as large as last year. Overall less than 10% of the jobs landed where we would like to see most of them, in the 60-100% segment. There's definitely room for improvement.

On the time consumed side, there's a clear trend even from 2018 of the 60-100% segment growing in size. Last year this segment presented about 25% of all consumed CPU time. Lets continue this trend in the future.

Other interesting observations

Long tail of users

During gathering statistics for an unrelated HPC usage query I noticed that we had over 320 individual users last year who submitted at least one job. However only three users submitted about 60% of all jobs while over 100 users submitted less than 1000 jobs. This gives us a typical long tail distribution, making any kind of fairshare implementation very difficult to impossible.