1 Estimate the performance of large scale run

We provide a model for estimating the performance of an execution with multiple file sets and multiple nodes based on that with single file set and one node. The same model can be applied to all steps of NGS Analyzer workflow: Alignment, Remove Duplicate and SNP. As the model does not include resource contention, calculated results are the maximum performance values.

Parameter	Type	Meaning
$Node\_Count$	Variable	Number of nodes (processes)
$File\_Count$	Variable	Number of target read files
$Elapse\_Time_1$	Constant	Elapse time of run with one file set and one node
$MFLOPS_1$	Constant	MFLOPS of run with one file set and one node
$MPIS_1$	Constant	MIPS of run with one file set and one node
$MEM\_Throughput_1$	Constant	Memory throughput of run with one file set and one node
$File\_Size_1$	Constant	Average file size of one file set

1.1 Elapse Time

$$\frac{File\_Count \times Elapse\_Time_1}{Node\_Count}$$

1.2 MFLOPS

$$Node\_Count \times MFLOPS_1$$

1.3 MFLOP

$$File\_Count \times Elapse\_Time_1 \times MFLOPS_1$$

1.4 MIPS

$$Node\_Count \times MPIS_1$$

1.5 Mega Instructions

$$File\_Count \times Elapse\_Time_1 \times MPIS_1$$

1.6 Memory Throughput

$$Node\_Count \times MEM\_Throughput_1$$

1.7 Memory Consumption

$$File\_Count \times Elapse\_Time_1 \times MEM\_Throughput_1$$

1.8 IO Read/Write Size

$$File\_Count \times File\_Size_1$$

1.9 IO Read/Write Throughput

$$\frac{Node\_Count \times File\_Size_1}{Elapse\_Time_1}$$

2 Estimate the required performance when maximum elapse time is given

We provide a model for estimating the required performance of each workflow step when the execution should be completed within the specified elapse time. The same model can be applied to all steps of NGS Analyzer workflow: Alignment, Remove Duplicate and SNP.

The meanings of parameters are same as the above table.

2.1 MFLOPS

$$\frac{File\_Count \times Elapse\_Time_1 \times MFLOPS_1}{Time}$$

2.2 MFLOP

$$File\_Count \times Elapse\_Time_1 \times MFLOPS_1$$

2.3 MIPS

$$\frac{File\_Count \times Elapse\_Time_1 \times MPIS_1}{Time}$$

2.4 Mega Instructions

$$File\_Count \times Elapse\_Time_1 \times MPIS_1$$

2.5 Memory Throughput

$$\frac{File\_Count \times Elapse\_Time_1 \times MEM\_Throughput_1}{Time}$$

2.6 Memory Consumption

$$File\_Count \times Elapse\_Time_1 \times MEM\_Throughput_1$$

2.7 IO Read/Write Size

$$File\_Count \times File\_Size_1$$

2.8 IO Read/Write Throughput

$$\frac{File\_Count \times File\_Size_1}{Time}$$

3 Estimate the required performance for staging when maximum staging time is given

We provide a model for estimating the required performance of stage-in/out of input/output data on a system that employs staging when the staging should be completed within the specified time. The same model can be applied to all steps of NGS Analyzer workflow: Alignment, Remove Duplicate and SNP.

Parameter	Type	Meaning
$Node\_Count$	Variable	Number of nodes where files are staged-in/out
$File\_Count$	Variable	Number of read files to be staged-in/out
$Time$	Variable	Maximum time for staging
DB_Size	Constant	Database size
$File\_Size_1$	Constant	Average file size of one file set

3.1 Stage In Throughput

$$\frac{Node\_Count \times DB\_Size + File\_Count \times File\_Size_1}{Time}$$

3.2 Stage Out Throughput

$$\frac{File\_Count \times File\_Size_1}{Time}$$