Многопоточная обработка данных для SPD

Что это, и почему получается так…

- Многотопоточная обработка данных характеризуется большим объёмом слабосвязанных задач, работающих параллельно на разнородных ресурсах, и обрабатывающих однородные данные большого объема
  - Интенсивно используется при обработке данных современных физических экспериментов
- На данный момент, переход на высокопроизводительную обработку в обозримом периоде не представляется возможным
  - нет прикладного ПО, для высокопроизводительных вычислительных систем (суперкомпьютеров): **тем не менее многопоточный и по возможности кроссплатформенный код будет необходим!**
  - существующая инфраструктура ориентирована на многопоточную обработку

Требования к организации данных для многопоточный обработки

- Данные собираются в файлы, файлы объединяются в наборы, наборы в коллекции
  - Файлы обрабатываются задачами и кроме специальных случаев, желательно что бы у задачи было небольшое число файлов.
  - Наборы обрабатываются заданиями. Грубо говоря - задание это набор однотипных задач, каждая из которых обрабатывает небольшой объем данных
  - Ветвь обработки - набор заданий которые необходимо осуществить в рамках одного процесса
- Размеры файлов разумные - не должны быть слишком маленькими в виду сложности контроля и учета, или же слишком большими - в виду затраты на загрузку, выгрузку таких файлов со счетных узлов
  - Здесь возможна горячая дискуссия о модных распределенных файловых системах… и как они не выдерживают необходимой нагрузки при массовых операциях ввода вывода
- Формат файлов должен подразумевать возможность разделения одного файла на части (split), или сборки одного файла из нескольких (merge)
  - Желательно, что бы формат был сопоставим с современными индустриальными стандартами для параллельного ввода/вывода (HDF5) ?

Размер (имеет значение)

- Все приведенные тут цифры приходят в основном из текущей практики.
  - Размерности рабочих файлов (файлов с данными) - от 4 до 10 Гб. Ограничения: скорость передачи, размер дисков на счетных узлах, размерность лент
  - Количество файлов в наборе: желательно до 100000. Нужно учитывать что в набор буду попадать не только файлы с данными, но и служебные файлы: логи задач, логи пилотов, отчеты о выполнении задач и т.п.
  - Наборы группируют файлы по определенным признакам
    - При этом один тот же файл может быть в разных наборах
    - Набор - единица репликации