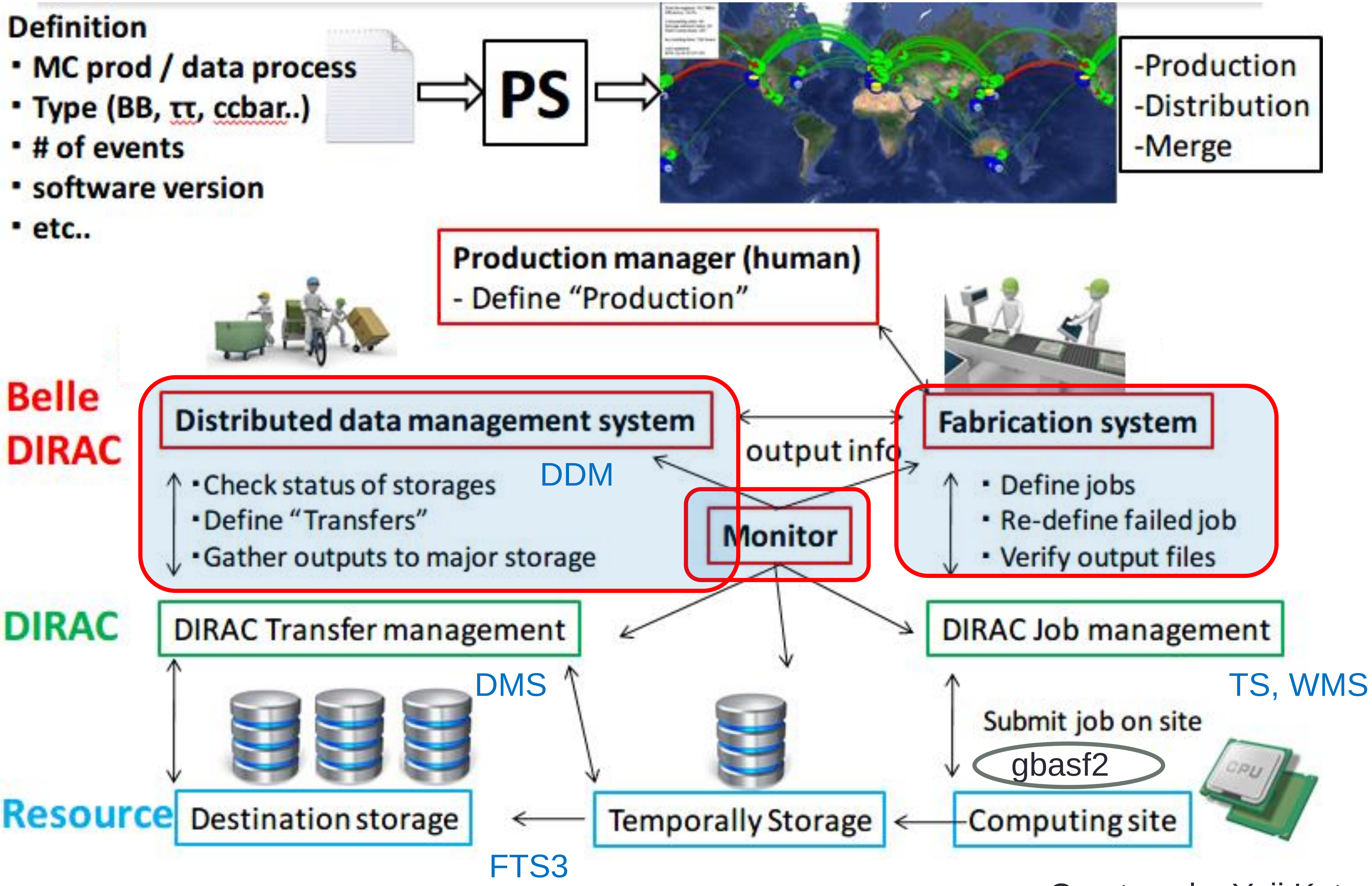
Belle II production system

Hideki Miyake (KEK)

2018 May 22nd, DIRAC users' workshop@Lyon

Belle II Production System



Definition
- MC prod / data process
- Type (BB, ττ, ccbar..)
- # of events
- software version
- etc..

PS

-Production
-Distribution
-Merge

Production manager (human)
- Define “Production”

Distributed data management system

DDM

- Check status of storages
- Define “Transfers”
- Gather outputs to major storage

output info

Monitor

Fabrication system

- Define jobs
- Re-define failed job
- Verify output files

DIRAC Transfer management

DMS

DIRAC Job management

TS, WMS

Submit job on site

gbasf2

Destination storage

Temporally Storage

Computing site

FTS3

Courtesy by Yuji Kato

Data Management Block

- A unit of Belle II data handling
  - All files stored on same SE
  - Dataset can consist of multiple DMBs (= different SEs)
- A DMB contains fixed number of files (say 1000 files)
  - If one file is unavailable by any reason, replaced by alternative
  - Job failure, SE down before transfer…

Fabrication System

Job goes to data location
No input data relocation for now

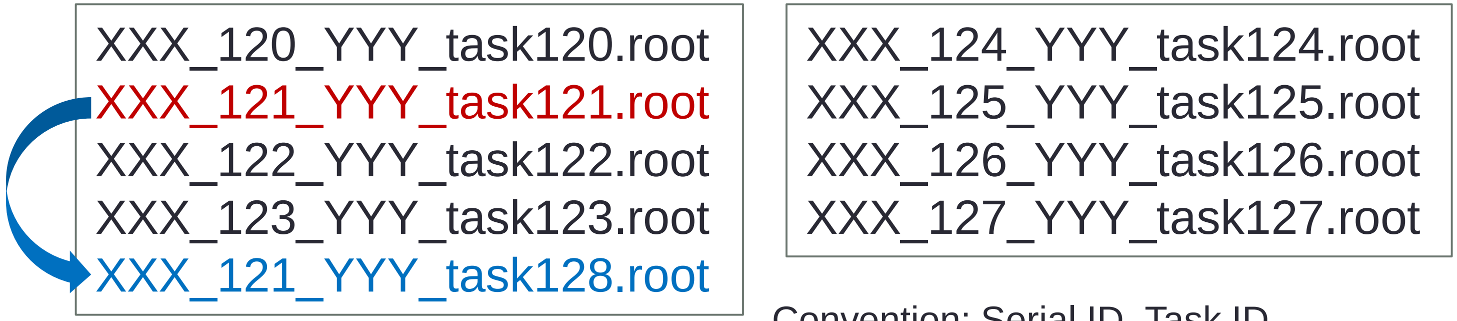
  - Each file is stored on temporary “local” SE → assembled by DDM

Distribution System (DDM)

Dataset: /xx/yy/BdecayA

/xx/yy/BdecayA/sub1

/xx/yy/BdecayA/sub2



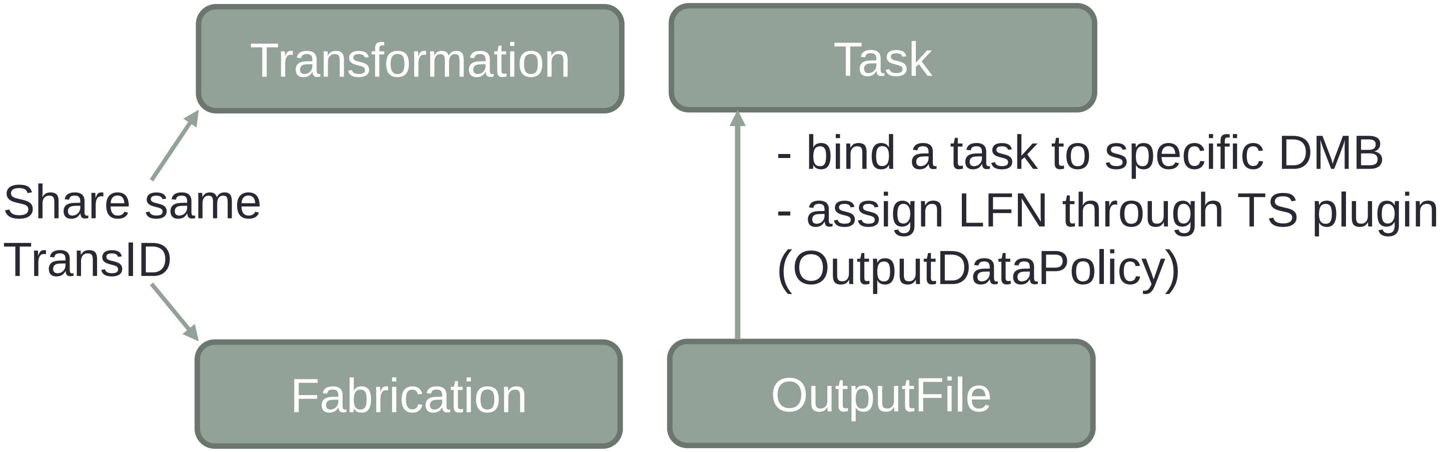
XXX_120_YYY_task120.root
XXX_121_YYY_task121.root
XXX_122_YYY_task122.root
XXX_123_YYY_task123.root
XXX_121_YYY_task128.root

XXX_124_YYY_task124.root
XXX_125_YYY_task125.root
XXX_126_YYY_task126.root
XXX_127_YYY_task127.root

Convention: Serial ID_Task ID

Fabrication System

- A kind of wrapper to DIRAC Transformation System
  - Provide DMB and output file management
  - A Fabrication is associated to specific Transformation
    - Practically make Fabrication instance with same Transformation ID
  - Designate output LFN when TS Task is initialized (“Created”→”Submitted”)



Transformation

Task

Share same
TransID

- bind a task to specific DMB
- assign LFN through TS plugin
(OutputDataPolicy)

Fabrication

OutputFile

Table structure: Fabrication
DataBlock
Outputfile

DMB1: /a/b/c/…/sub00/xxx1~1000.root
DMB2: /a/b/c/…/sub01/xxx1~1000.root

Fabrication Agent

- Take over the role of ValidateOutputDataAgent
  - Validate file metadata (checksum of SE and LFC entry)
  - Wait for processing if any RMS request is open

- Ask DDM to transfer output file
  - Watch transfer status

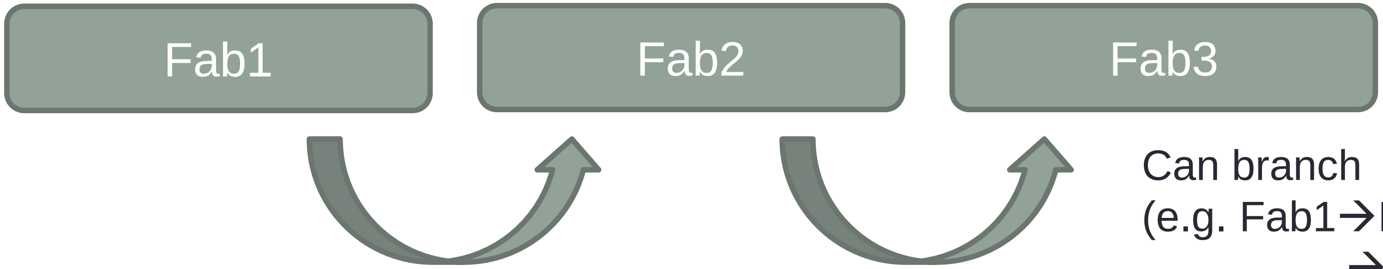
- Failure recovery:
  - Task failure
    - Failed task is not rescheduled but replaced by new Task (new LFN)
    - Release assigned input files (
    - Check “removed” but used as input files too
  - Transfer failure
    - Drop the Task and generate new

- Manage data block
  - Remove unnecessary files (corrupted, wrong production definition…)
  - Fill datablock/dataset metadata
  - Fix DB inconsistency (including file status of TransformationFiles)

Production Management System

- Belle II PMS manages various tasks
  - Generate/monitor Transformation and Fabrication instances based on production request (written in json)

  - Chain output files generated by former Transformation to next (take over InputDataAgent…but not by metadata but by timestamp)
    - After data transfer by DDM (thus latter TS Task runs on limited number of sites)



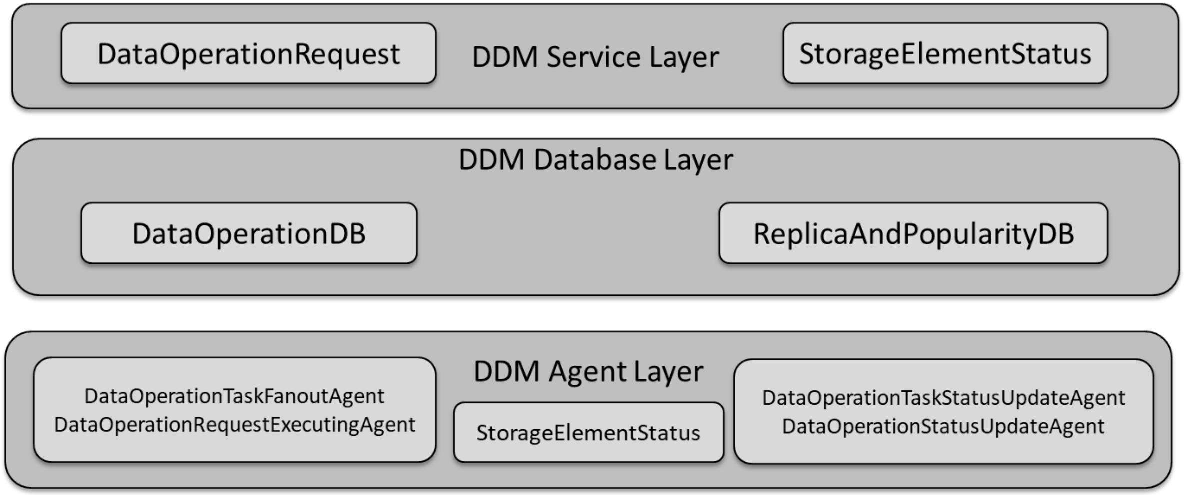
Can branch
(e.g. Fab1→Fab2→Fab3
→Fab2’→Fab3’)

  - Verify production which consumed all input files
    - Check consistency among input and output data files (e.g. not doubly used)

Automatize misc management tasks: flush, additional WMS priority control (long waiting job or very last Tasks in TS), diagnosis/fix of stuck production

Distributed Data Management System

- Data transfer not using TS but FTS (through RMS) directly
  - Optimized for data management block scheme
- Monitor each SE status stored in own DB
- Automatically determine destination SE by predefined policies (e.g. by free space)
- Bulk deletion (doesn’t make stress on both LFC and each SE)
- Coordinate transfer and deletion requests to avoid race condition



Extension to existing components

- Transformation System
  - MCExtensionAgent
    - Controlled by total Waiting (and submitting) jobs for specific JobType
    - Consider priority
  - TaskManager
    - Doesn’t back “Reserved” Task status to “Created” but “Failed”
      - Since “Created” Task repeats task initialization (e.g. LFN assignment)
  - TransformationPlugin
    - Our own logic to control Task creation
    - Doesn’t submit new Task if submitting Task exists in the Transformation or a datablock has sufficient number of submitted Task

Our model is to submit new Tasks gradually even if tons of input data is given

- Core
  - OutputDataPolicy
- WorkloadManagement System
  - Executor (InputData and JobScheduling)
    - Skip staging check… some our inputdata files are distributed to ~20 SEs and don’t want to check all replicas per WMS job submission

What can be common?

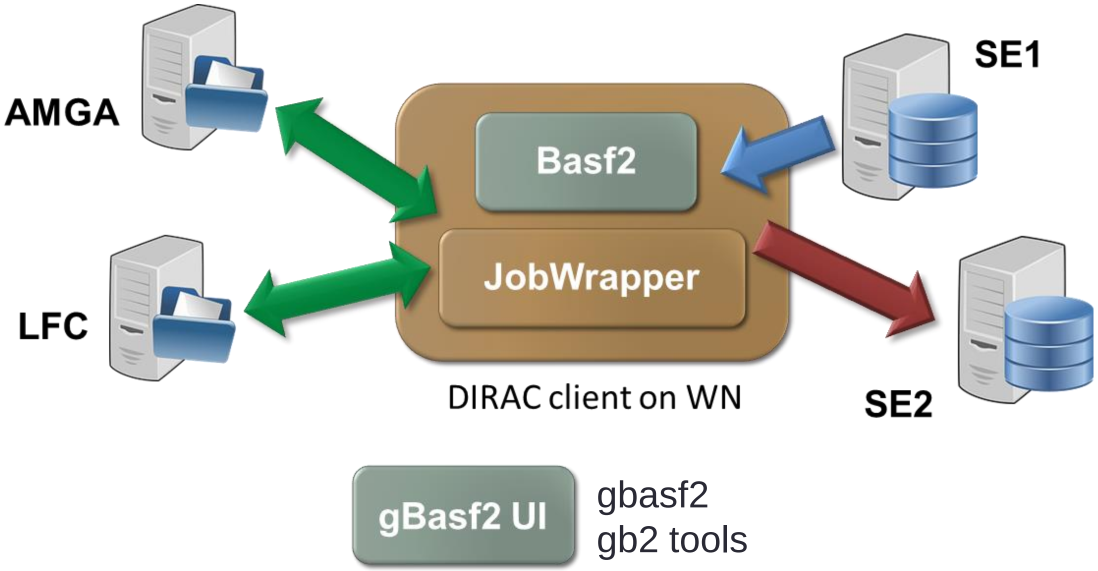
- Belle II production system:
  - ProductionManagement
  - Fabrication
  - DistributedDataManagement

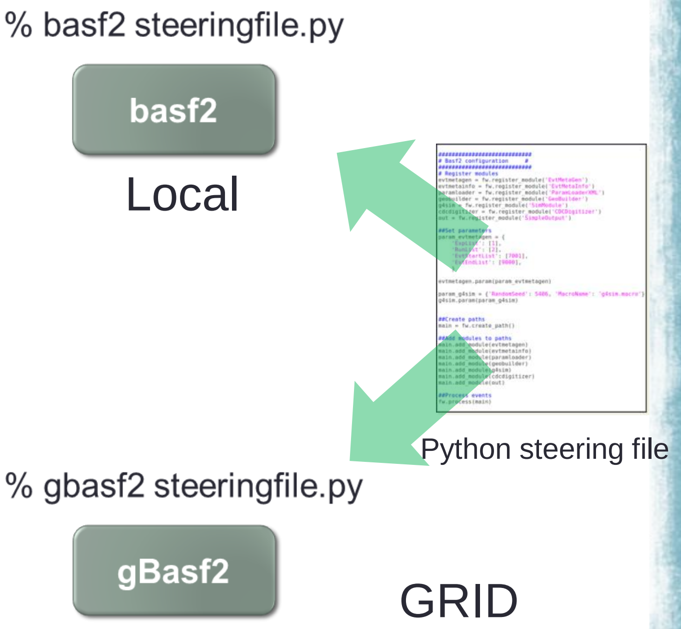
- Oppositely…
  - Common production system usage in Belle II?
    - Massive but simple job workflow (e.g. user analysis)

Backup

gBasf2

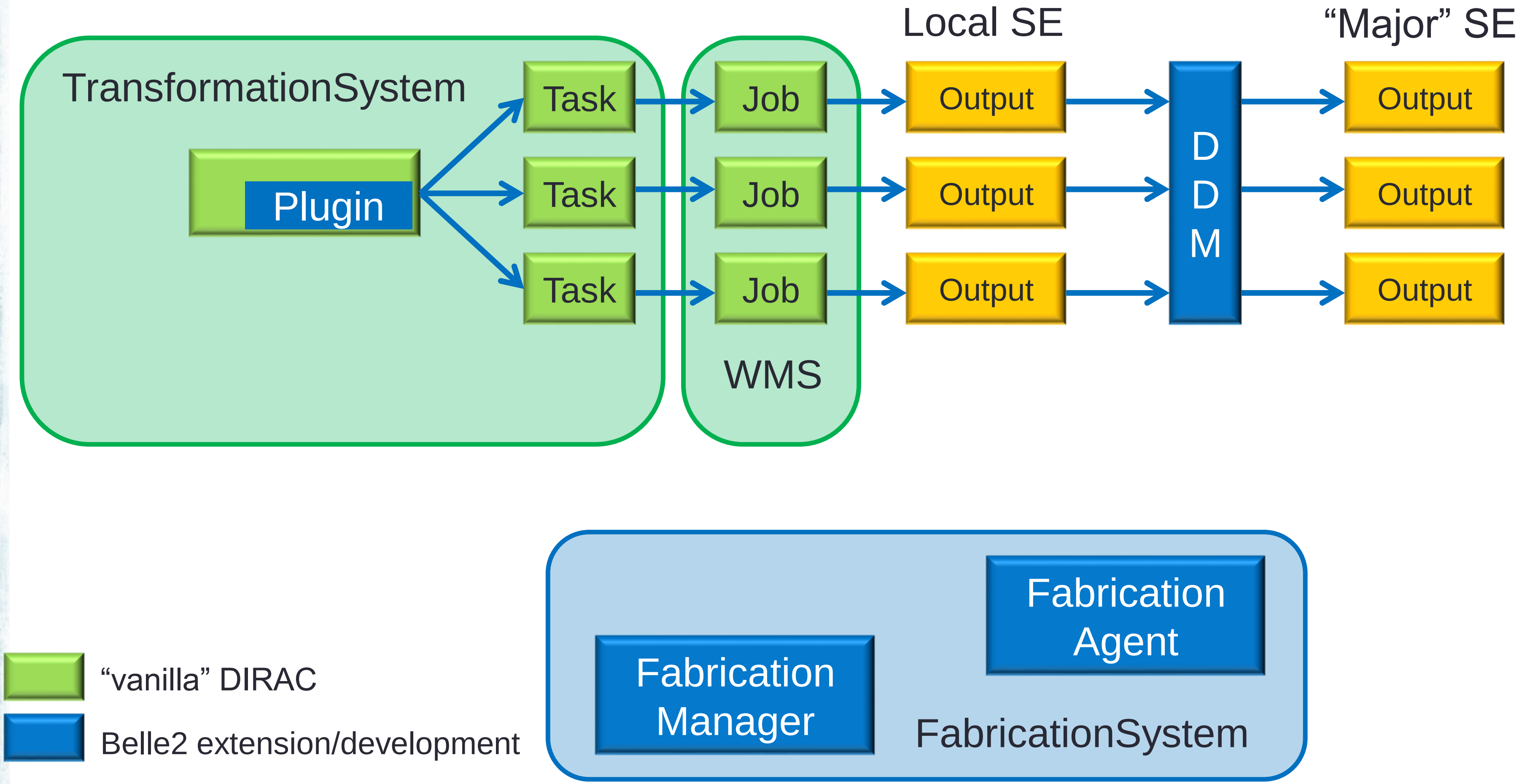
- Basf2 is our analysis software framework
  - Modular basis job executing platform
- Interface to distributed computing is given by gBasf2 (grid Basf2)
  - Provide transparent job execution on DC environment
    - Data input/output, file catalog/metadata registration…
  - Provide also collection of tools to handle job and data through DIRAC API (gb2 tools)





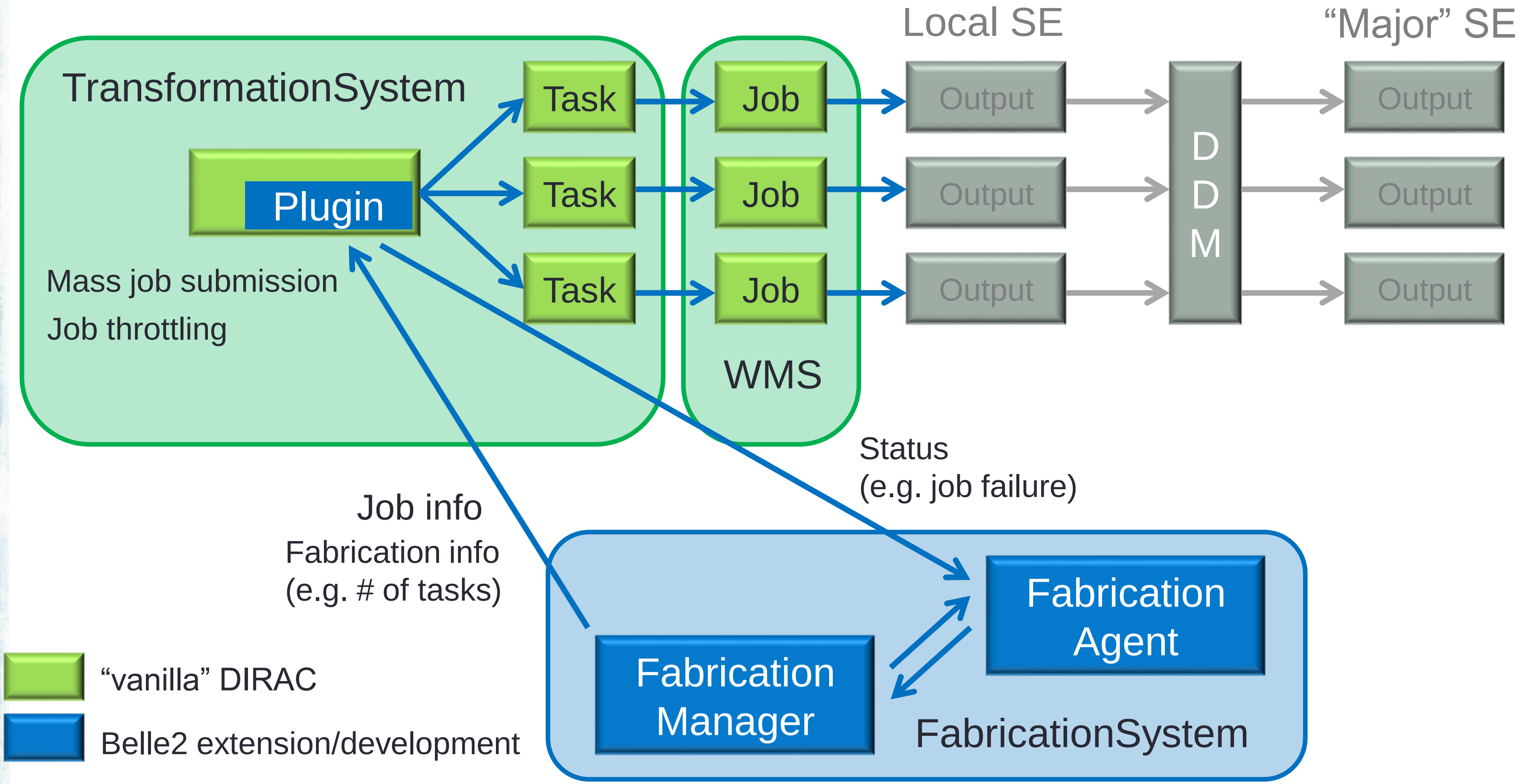
Workflow: overview

- Fabrication System exploits existing DIRAC components; TransformationSystem (TS) and WorkloadManagementSystem (WMS)
- TS is controlled by our plugin extension



Workflow: job management

- FS controls both job submission and failure job resubmission
- Each job status is monitored through TS



TransformationSystem

Plugin

Mass job submission

Job throttling

Task

Task

Task

Job

Job

Job

WMS

Local SE

Output

Output

Output

DDM

"Major" SE

Output

Output

Output

Status
(e.g. job failure)

Job info

Fabrication info
(e.g. # of tasks)

Fabrication Agent

Fabrication Manager

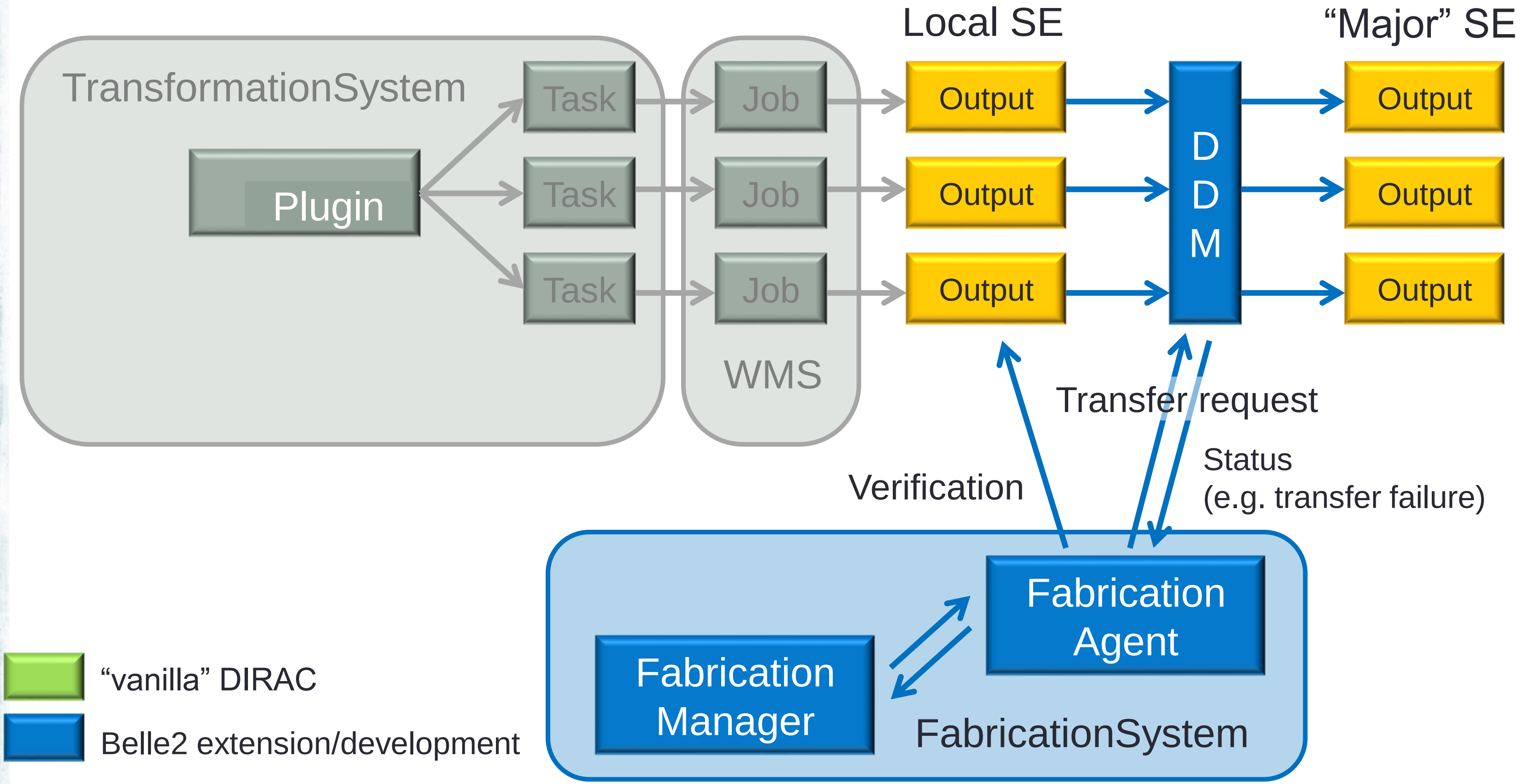
FabricationSystem

🟩 "vanilla" DIRAC

🟦 Belle2 extension/development

Workflow: file management

- Once output file is created by GRID job, verifies each
- If file status is good, ask for DDM to transfer the file


Local SE
"Major" SE
TransformationSystem
Plugin
Task
Task
Task
Job
Job
Job
WMS
Output
Output
Output
D D M
Output
Output
Output
Transfer request
Status
(e.g. transfer failure)
Verification
Fabrication Agent
Fabrication Manager
FabricationSystem
"vanilla" DIRAC
Belle2 extension/development