University of North Carolina at Chapel Hill - High Volume of Concurrent Ingests

Title (goal)	High Volume of Concurrent Ingests
Primary Actor	Submitter
Scope	
Level	
Story	We need to be able to reliably load submission packages on a large scale from local network drives. We ingest many thousands of objects as part of a single batch submission. We would like to be able to ingest objects in the batch in parallel for higher throughput. The batches will **not** be part of a single Fedora transaction with rollback. (Our tools will sometimes pause, re-prioritize and then resume a batch ingest job.) To support ongoing collection work an individual batch submission should take no longer than 2 days to ingest, given sufficient i/o and cluster resources. We anticipate approximately 10 submissions per month, each having 10k items and totaling 1TB. So that leaves a **base ingest load of 100k items at 10TB per month**, where *each item is an object plus one primary data stream and around 4 smaller streams*. **Scaling out to meet occasional load:** Our largest anticipated collection next year is a 10TB collection containing many video files. That would come in the form of perhaps 10 submission packages of 1TB each, containing files numbering in the hundreds. We'd like to be able to scale out our cluster to handle such an oversized collection without disrupting base ingest load. That leaves a maximum ingest load of **200k items totaling 20TB per month**.
Notes	Individual files in the video collection example will include HD video files that are several hours in length. So there will be files that extend to approx. 250GB (rough estimate of 4hrs JP2 at 1080p, 8bit color, 25fps)