Novel semantics of fault tolerant collective communication operations

Communication operations among a potentially large set of (possibly remote) processes are called collective communication operations. Examples include broadcast, reduce, and allreduce. Many more operations are defined in common communication libraries, e.g., MPI.

Fail-stop failures of participating processes lead to the respective process not contributing. Since these communication operations are typically implemented by repeatedly having pairs of processes send messages, a failed process can also impede the propagation of the contribution of other processes. This happens, e.g., when a failed process is sent a message and is expected to forward the message to other processes.

Functions that implement a collective communication operation have multiple options for the semantics with respect to failed participants. This talk describes the semantics that User Level Failure Mitigation (ULFM) adds to MPI as well as new semantics and discusses the differences. New semantics for some operations provide consensus. Unlike, e.g., Paxos, this talk targets large sets of processes.

The talk raises the question – but does not answer it – which semantics are most useful for applications and communication libraries. One goal of the talk is to get feedback on what semantics should be available.