Lots of PoS based BFT algorithms are claiming to have very high rates of throughput*, combined with low latency. Much more so than what one finds in deployed Nakamoto consensus chains.

What critical tradeoff or assumption is made in these algorithms which unlocks this capacity improvement?

*I mean high rates on the base layer, not about enabling sharding or any L2 solution.