In modern compute centers, there are thousands of processors. In our vision of On-The-Fly compute centers the number of processors in a compute center presumably will be even more. These processors (or subsets of them) often share a common resource such as memory or data connection. Sometimes, this common resource can become the bottleneck of the whole system, especially if data throughput is high.

Now, OTF services are typically composed of different smaller services in a certain manner. In our theoretical models, we take these compositions into account by considering precedence constraints.

As a demonstration of our techniques and ongoing research, I will introduce a theoretical model taking into account both properties by considering a multiprocessor environment with identical, fixed-speed processors sharing a common resource or bandwidth. For this model, we assume the precedence constraints to be linear, that is every job has at most one predecessor and one successor. The assignment of the services to the processors is fixed, i. e. there is one queue of jobs on every processor and they have to be finished in that order. The bandwidth is the bottleneck of the system, i. e. if a job has to be slowed down, this is due to the required bandwidth not provided in some time step (we consider the problem as a discrete model). Processing only parts of a job in one time step is allowed - this is the new aspect in comparison to previous theoretical models. I present basic strategies and ideas to solve the problem and future modifications that could be interesting.