Purpose: To reconstruct in real-time the dose distribution of intensity-modulated radiation therapy (IMRT) delivered to a patient during treatment. **Method and Materials:** First, we use a GPU-based FSPB algorithm to calculate the dose deposition coefficients *Dij* for all IMRT field. *Dij* is the dose contribution of the *i*th beamlets to the *j*th voxel on the patient image, and it is pre-computed and stored in the memory. During the treatment delivery, MLC leaf positions will be read in real-time from either the MLC controller or the cine EPID images. Then the delivered dose distribution is generated using the *Dij* matrix and the leaf positions. In this work, we simulate this process using the MLC DynaLog files in retrospective. We extract the leaf positions and fractional MU every 50 ms from the MLC DynaLog files and converted them to a beam mask *Bi* using an in-house software based on MATLAB. The accumulated dose distribution delivered up to a time point is then computed by convolving *Dij* with *Bi*. **Results:** It takes on average less than 1 s to pre-calculate *Dij* for each beam. It takes on average 0.2 s to update the accumulative dose distribution on patient's geometry. It is expected that this time can be further reduced by 10 to 20 times when we implement it on GPUs to meet the real-time requirement. **Conclusion:** We have developed a dose reconstruction algorithm with pre-calculated dose deposition coefficients using GPU-based FSPB algorithm. The results indicate that real-time dose reconstruction during treatment delivery is feasible, enabling dose-based real-time treatment monitoring.