**Purpose:** Cone-beam CT (CBCT) plays an important role in image guided radiation therapy (IGRT). However, the large radiation dose from serial CBCT scans in most IGRT procedures raises a clinical concern. This excessive imaging dose can be reduced by reducing the number of x-ray projections and/or lowering mAs levels per projection. It is the goal of this work to develop a fast GPU-based algorithm to reconstruct high quality CBCT from undersampled and noisy projection data so as to lower the imaging dose.

**Method and Materials:** The CBCT is reconstructed by minimizing an energy functional consisting of a data fidelity term and a total variation regularization term. We developed a GPU-friendly version of the forward-backward splitting algorithm to solve this model. Multi-grid technique is also employed.

**Results:** It is found that 20–40 x-ray projections are sufficient to reconstruct images with satisfactory quality for IGRT. The reconstruction time ranges from 77 to 130 sec on an NVIDIA Tesla C1060 GPU card, depending on the number of projections used, which is estimated ~100 times faster than similar iterative reconstruction approaches. Phantom studies indicate that our algorithm enables the CBCT to be reconstructed under as low as 0.1 mAs/projection level. Comparing with currently widely used full-fan head-and-neck scanning protocol of ~360 projections with 0.4 mAs/projection, it is estimated that an overall 36–72 times dose reduction has been achieved. Our algorithm can also be used in CBCT reconstructions in limited field of view scanning protocols for further reducing dose and in limited angle scanning protocols for shortening scanning time.

**Conclusion:** The developed GPU-based CBCT reconstruction algorithm is capable of lowering imaging dose considerably. The high computational efficiency makes this iterative CBCT reconstruction approach applicable in real clinical environments for the first time.